FITAT/ISPM 2014

http://www.fitat.org

ISSN : 2288-9973

Join us July 29 - August 1
Chiang Mai, Thailand

General Chairs

Keun Ho Ryu, Chungbuk National University, South Korea
Nipon Theera-Umpon, Chiang Mai University, Thailand
Nyambaa Shirnen, National University of Mongolia, Mongolia

PC Chairs

Oyun-Erdene Namsrai, National University of Mongolia, Mongolia
Goutam Chakraborty, Iwate Prefectural University, Japan
Sansanee Auephanwiriyakul, Chiang Mai University, Thailand

Keynote Speakers

Sermkiat Jomjunyong, Chiang Mai University, Thailand
Eun Jong Cha, Korean NRF, South Korea
Goutam Chakraborty, Iwate Prefectural University, Japan
Anan Phonphoem, Kasetsart University, Thailand
Sansanee Auephanwiriyakul, Chiang Mai University, Thailand

Organized by:

TCEB
THAILAND CONVENTION & EXHIBITION BUREAU

МОНГОЛ УЛСЫН ИХ СУРГУУЛЬ
Үнэн мэдлэг, оюуныг дээдэлнэ

Dear Distinguished Delegates and Guests,

On behalf of the FITAT 2013/ISPM 2014 Conference Committee, it is our great pleasure to welcome you to Chiang Mai, Thailand for the 7th International Conference on Frontiers of Information Technology, Applications and Tools, and the 4th PT-ERC International Symposium on Personalized Medicine. We are grateful to have this year's conference held in Chiang Mai, Thailand. The International Conference on FITAT 2014 and International Symposium on ISPM 2014 are prominent international forums providing a platform for presentations and discussions of recent developments and future trends in Information Technology and Personalized Medicine. We hope that you will benefit both scientifically and personally from the forum and enjoy the FITAT 2014/ISPM 2014.

This volume contains the Proceedings of the FITAT 2014/ISPM 2014, held in Chiang Mai, Thailand, July 29 – August 1, 2014. The FITAT 2014/ISPM 2014 conference includes 5 keynote speeches, 30 papers in oral presentation where each paper is assigned into one of the 10 conference sessions, and 37 papers in interactive presentation session.

The papers were submitted from many countries across the globe and many program committee members from 6 countries worked hard to prepare the FITAT 2014 and ISPM 2014. We, on behalf of the Program Committee of FITAT 2014, 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
July 29 – August 1, 2014
Chiang Mai, Thailand

PROGRAM OF FITAT/ISPM 2014

Outline

Day 1 ISPM Poster and Discussion: 13:00 ~ 18:00, Tuesday, July 39, 2014

Day 2 FITAT/ISPM Oral and Poster sessions: 08:30 ~ 18:30, Wednesday, July 30, 2014
Banquet and Awarding Ceremony: 18:30 ~ 21:00, Wednesday, July 30, 2014

Day 3 FITAT/ISPM Oral and Poster sessions: 08:30 ~ 18:00, Thursday, July 31, 2014

Day 4 FITAT Business meeting: 08:30 ~ 12:00, Friday, August 1, 2014

Tuesday, July 29, 2014

ISPM Poster and Panel discussion (13:00 ~ 18:00)

Time	Sessions
13:00 ~ 15:00	**Registration** **Location: Lobby 1**
15:00 ~ 18:00	**Poster and Discussion** **Location: Lobby 1** Session chair: Kyung-Ah Kim and Ho Sun Shon **P1-01: PE malDetector: An Intelligent Malware Detection System Based on Learning Algorithms** *Munkhtuya.D, Usukhbayar.B, Uitumen.J, Sodbileg.Sh, Nyamjav.J* **P1-02: TCP Based Attack Detection Using WEKA Tool** *Ugtakhbayar.N, Enkhjin.B, Sodbileg.Sh* **P1-03: A Prototype of Expert System for Rural Medical Centers** *Uranchimeg Tudevdagva, Uyanga Sambuu, Yumchmaa Ayush* **P1-04: Mongolian Traditional Stamp Recognition** *Gantuya Perenleilhundev, Suvdaa Batsuuri* **P1-05: Adding Mongolian Language Module to the Asterisk Voicemail** *Jamiyan Sukhbaatar, Nyamjav Jambaljav, Bat-Erdene Batdolgor* **P1-06: Real-time Hand Gesture Recognition using SVM** *Suvdaa Batsuuri, Chintogtokh Batbold* **P1-07: OOD Metrics for Cohesion - A Survey** *Batnyam Batttulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar* **P1-08: Implementation of XMI Parser** *Batnyam Batttulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar* **P1-09: Data Mining Techniques used to Improve Sport Prediction** *Tsend-Ayush Sh, Otgonnaran O, Oyun Erdene Namsrai* **P1-10: Performance Improvement of Mining Techniques: Supermarket's Data Analysis** *Tsatsral Amarbayasgalan, Bilguun Jargalsaikhan, Otgonnaran O, Oyun-Erdene Namsrai*

Wednesday, July 30, 2014

FITAT/ISPM Oral Presentations (08:30 ~ 18:00)

Time	Sessions
08:30 ~ 09:00	**Registration** **Location: Lobby 1**
09:00 ~ 10:10	**Session Name: FITAT/ISPM Opening and Keynote Session 1** **Location: Room 1** Session Chair: Oyun-Erdene Namsrai **Opening speech (10min)** Keun Ho Ryu (Chungbuk National University) Nipon Theera-Umpon (Chiang Mai University) Nyambaa Shirnen (National University of Mongolia) **Understanding EEG signal for Better Brain-Computer Interface and other Applications (30min)** Goutam Chakraborty *Iwate Prefectural University, Japan* **Fuzzy Set Theory in Information Technology (30min)** Sansanee Auephanwiriyakul *IEEE Senior Member, Head of Computer Engineering Department, Chiang Mai University, Thailand*
10:10 ~ 10:30	**Coffee Break**
10:30 ~ 12:00	**Session Name: Core Database Technologies and Data Mining Techniques** **Location: Room 1** Session Chair: Musa Ibrahim M. Ishag **Pattern Mining from Online Social Media (30min)** Basabi Chakraborty, Takako Hashimoto *Iwate Prefectural University, Japan* **Grid-based Image Morphing (20min)** Porawat Visutsak *King Mongkut's University of Technology North Bangkok, Thailand* **Adeptness Associative Learning Method for Real-Time Cardiac Arrhythmia Detection (20min)** Mohamed Ezzeldin A. Bashir, Dong Gyu Lee, Ibrahim Musa Ishaq, Makki Okasha, Ho Sun Shon , Keun Ho Ryu *University of Medical Science and Technology, Sudan* **A Novel Mathematical Descriptive System for Human Body-Shape Representation (20min)** Sukationg Phuphatana, Pirawat WATANAPONGSE *Department of Computer Engineering, Kasetsart University, Thailand*
12:00 ~ 13:00	**Lunch** **Location: Restaurant 1**

13:00 ~ 14:40	**Session Name: Communication and Networking** **Location: Room 1** Session Chair: Seon-Phil Jeong **Three-dimensional Image Processing using Integral Imaging (30min)** Ganbat Baasantseren *National University of Mongolia, Mongolia* **Mining Frequent Itemsets in Transactional Database by Reduction of Trees Traversal (20min)** Supatra Sahaphong, Gumpon Sritanratana *Ramkhamhaeng University, Thailand* **Assessment of E-learning Readiness in National University of Mongolia (20min)** Otgontsetseg Sukhbaatar, Tsolmon Zundui, Lodoiravsal Choimaa *National University of Mongolia, Mongolia* **Extracting Political Networks of the Sudan from Online Newspapers (20min)** Musa Ibrahim M. Ishag, Ho Sun Shon, Keun Ho Ryu *Chungbuk National University, South Korea*
14:40 ~ 15:00	**Coffee Break**
15:00 ~ 16:20	**Session Name: Data Mining Applications** **Location: Room 1** Session Chair: Wang Ling **Fuzzy Measure Application to Decision Making (30min)** Sanghyuk Lee *Xi'an Jiaotong-Liverpool University, China* **Real-Time Data Warehousing and Online Analytical Mining of Re-designed Large Database: Challenges and Solutions (30min)** Oyun-Erdene Namsrai *National University of Mongolia, Mongolia* **Differential Wheeled Mobile Robot Self-localization Method for 8 Bit Microcontroller (20min)** Batbayar Unursaikhan, O.Zoljargal *National University of Mongolia, Mongolia*
16:20 ~ 18:00	**Session Name: Poster and Discussion** **Location: Lobby 1**
18:00 ~ 21:00	**Session Name: Special Keynotes and Banquet** **Location: Restaurant 1** Session Chair: Keun Ho Ryu and Nipon Theera-Umpon **Triple Helix Enhancing Innovation (15min)** Sermkiat Jomjunyong *Chiang Mai University, Vice President for Research and Academic Services, Thailand* **Policies for Promoting Basic Convergence Research by the National Research Foundation(NRF) of Korea (15min)** Cha Eun Jong *Korean NRF, South Korea* **Banquet and Awarding Ceremony**

Wednesday, July 30, 2014

FITAT Poster Presentation (13:00 ~ 18:00)

Time	Sessions
13:00 ~ 18:00	**Session Name: Interactive Session 2** **Location: Lobby 1** Session Chair: Ho Sun Shon **P1-01: Comparison of Prognosis Factors between ST-Segment Elevation Myocardial Infarction and non-ST-Segment Elevation Myocardial Infarction of Patients with Atrial Fibrillation** *Ho Sun Shon, Jang-Whan Bae, Byung Jun Cho, Young Sung Lee, Young Gyu Kim* **P1-02: Short-Term Electricity Price Forecasting using Cascade Neural Network** *Cheng Hao Jin, Hyun Woo Park, Ling Wang, Kyung Hee Lee* **P1-03: Ensemble Method based MicroRNA Selection for Disease Diagnosis** *Minghao Piao, Yongjun Piao, Feifei Li, Keun Ho Ryu* **P1-04: The Construction of Integration Dataset for Correlation Analysis of Heart disease and Meteorological Information** *Hyeongsoo Kim, Kwang Sun Ryu, Jae Won Lee, Kwan Hee Yoo* **P1-05: The Generation of Fusion factor for Acute Myocardial Infarction based on Causal Association Rule Mining** *Kwang Sun Ryu, Seung Hyeon Yang, Hyun yoo Park, Soo Ho Park, Ibrahim M. Ishag, Jang Whan Bae* **P1-06: Biomedical Event Extraction with Random Forests** *Tsendsuren Munkhdalai, Meijing Li, Khuyagbaatar Batsuren, Wan-Sup Cho* **P1-07: CUDA-based Multiple Linear Regression for Analysis of Large Health Data** *Soo Ho Park, Ho Sun Shon, Eun Jong Cha* **P1-08: A Progressive Architecture for Source Code Clone Detection and Extraction by Using Data Ming Methods and MapReduce Paradigm** *Dingkun Li, Minghao Piao, Jong Yun Lee* **P1-09: Correlation analysis of RNA-Seq and qRT-PCR for detection of differentially expressed genes** *Yongjun Piao, Nak Hyeon Choi, Meijing Li, Keun Ho Ryu* **P1-10: Gait identification using Wearable Sensors with Spatio-Temporal Features** *Hyun Woo Park, Kyeong Seok Lee, Soo Ho Park, Cheng Hao Jin* **P1-11: Development of System to Measure Length of Moving Object on CCTV Image** *Kyu Ik Kim, Sunny Song, Myung-Sic Kim, Jin Suk Kim* **P1-12: Accelerating the morphology operations using a CUDA on Graphics Processing Units** *Amartuvshin Renchin-Ochir, Gerelttulga Galaa, Bolormaa Dalanbayar* **P1-13: Programmable logic chip based Sinewave generator using Cordic algorithm** *Battogtokh.J, Zorig.B, Bolormaa.D* **P1-14: Technical analysis on 3D reconstruction methods** *Erdenebat Budsuren, Baatarbileg Altangerel, Oyun-erdene Namsrai*

Thursday, July 31, 2014

FITAT/ISPM Oral Presentation 08:30 ~ 18:00

Time	Sessions
08:30 ~ 09:00	**Registration** **Location: Lobby 1**
09:00 ~ 10:10	**Session Name: Keynote Session 3** **Location: Room 1** Session Chair: Sansanee Auephanwiriyakul **Low-Cost Wireless Sensor and Communication System for Landslide Monitoring and Assessment (30min)** Anan Phonphoem *Kasetsart University, Head of Computer Engineering Department, Thailand* **Combining Tag and Value Similarity for Data Extraction and Alignment (30min)** Weifeng Su *BNU-HKBU United International College, China*
10:10 ~ 10:30	**Coffee Break**
10:30 ~ 12:00	**Session Name: Biomedical Informatics** **Location: Room 1** Session Chair: Minghao Piao **Practical Problems of Developing Indoor Positioning Systems using WiFi (30min)** Tung Hoang Do Thanh, Tien Nguyen Ba, Binh Ngo Van *Head of Management Systems Department, Vietnam Institute of Information Technology (IOIT) of Vietnamese Academy of Science and Technology (VAST), Vietnam* **Diagnosing patient with Acute Myocardial Infarction using mRNA Profiles (20min)** Seung Hyeon Yang, Kwang Sun Ryu, Musa Ibrahim M. Ishag *Chungbuk National University, South Korea* **Exploration of MicroRNA-Based Cancer Classification Using Decision Tree Classifier (20min)** Feifei Li, Minghao Piao, Keun Ho Ryu *Chungbuk National University, South Korea* **Comparison of Combination of Feature Selection Methods and Classification Methods for Multiclass Cancer Classification from RNA-seq (20min)** Nak Hyeon Choi, Yongjun Piao, Meijing Li, Keun Ho Ryu *Chungbuk National University, South Korea*
12:00 ~ 13:00	**Lunch** **Location: Restaurant 1**

13:00 ~ 14:50	**Session Name: PSM Session** **Location: Room 1** Session Chair: Hoang Do Thanh Tung **Shape Representation Using Morphological Granulometries (30min)** Nipon Theera-Umpon *IEEE Senior Member, Director of Biomedical Engineering Center, Chiang Mai University, Thialand* **A Personalized u-commerce Recommender System using Bayesian Learning and Weighted Preference (20min)** Seon-Phil Sunny Jeong *BNU-HKBU United International College, China* **Distance Metric Learning for Face Recognition (20min)** Suvdaa Batsuuri *National University of Mongolia, Mongolia* **Design for Triple Helix model framework using information of bibliography (20min)** Gwi Suk Gim, Ho Sun Shon, Byung Jun Cho, Hyung Chul Rah, So Young Kim *Chungbuk National University, South Korea* **Big Data based Framework Design for Korean Patients with Acute Myocardial Infarction (20min)** Changwoo Woo, Wooyeong Jang, Ho Sun Shon, Eung-Do Kim, Gilwon Kang *Chungbuk National University, South Korea*
14:50 ~ 15:10	**Coffee Break**
15:10 ~ 17:00	**Session Name: PSM, Text Mining and Natural Language Processing** **Location: Room 1** Session Chair: Otgontsetseg Sukhbaatar **Invited speech (30min)** Wang Ling *The North East Dianli University, China* **Keyword Extraction using Anti-pattern (20min)** Khuyagbaatar Batsuren, Tsendsuren Munkhdalai, Meijing Li, YoungJung Kim, JongYun Lee *Chungbuk National University, South Korea* **Developing Graph Database for Multilingual Corpus (20min)** Hyeon Ah Park, Khuyagbaatar Batsuren, Nak Hyeon Choi, Jeong Hee Hwang *Chungbuk National University, South Korea* **Classification of Diseases from Number of Outbreaks (20min)** Wooyeong Jang, Changwoo Woo, Ho Sun Shon, Young-Sung Lee, YoungGyu Kim, Keun Ho Ryu *Chungbuk National University, South Korea* **Development of Web-based System for Analysis of Urinary Cancer Patient from Disease Prevention Questionnaire (20min)** Kyeong Seok Lee, Hyun Woo Park, Soo Ho Park, Kyung Ah Kim *Chungbuk National University, South Korea*

17:00 ~ 17:20	**Coffee Break**
17:20 ~ 18:00	**Session Name: Communication and Signal Processing** **Location: Room 1** Session Chair: Suvdaa Batsuuri **Determination of Surface Radio Refractivity over Mongolia (20min)** Jamiyan Sukhbaatar, Nyamjav Jambaljav, Damdinsuren Erkhembayar *National University of Mongolia, Mongolia* **Constructing a System for Monitoring, Managing Groundwater in the Industrial Zones of Hanoi City (20min)** Vu Thi Hong Nhan *Vietnam National University, Vietnam* **Investigation of SEE on a 32-bit Microprocessor based on SPARC V8 Architecture by Laser Test (20min)** Chunqing Yu, Long Fan, Suge Yue, Maoxin Chen, Shougang Du *Beijing Microelectronics Technology Institute, China*

Thursday, July 31, 2014

ISPM Poster Presentation (13:00 ~ 18:00)

Time	Sessions
13:00 ~ 18:00	**Session Name: Interactive Session 3** **Location: Lobby 1** Session Chair: Mi Sug Gu and Cheng Hao Jin **P1-01: Integrated Public Bike Rental System Design** *Sunny Song, Kyu Ik Kim, Myung-Sic Kim, Keun Ho Ryu* **P1-02: Study on the Method of Composing Data Warehouse for Error Verification and Multi-dimensional Error Test** *Myung-Sic Kim, Gwi-Seop Song, Kenu Ho Ryu* **P1-03: Wireless Uroflowmetry System for Self-test at Home** *In-Kwang Lee, A-Rong Heo, Ho Sun Shon, Keun Ho Ryu, Kyoung-Ok Kim, Eun-Jong Cha, Kyung-Ah Kim* **P1-04: Context Ontology based Mobile Information Retrieval** *Mi Sug Gu, Ho Sun Shon, Keun Ho Ryu* **P1-05: Protected Health Information for Research on Computer Forensics** *Yoonhwan Shin, Keun Ho Ryu* **P1-06: Evaluating the Impact of Design Patterns on Code Design using Object-Oriented Metrics** *Batnyam Battulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar* **P1-07: 3D Reconstruction from Uncalibrated Images** *Tsetsegjargal Erdenebaatar, Suvdaa Batsuuri*

	P1-08: Polynomial Approximation of Impedance of Microstrip Patch Antenna *Batpurev Mongol, Gerelmaa Byambatsogt, Ganbat Baasantseren* **P1-09: Hand Gesture Controlled Drawing Tool using "Asus xtion pro"** *Amartuvshin Renchin-Ochir, Dorjnamjirmaa Badraa* **P1-10: Analyzes of Enrollment Database of a University Information System** *Bulganchimeg.B, Naranchimeg.B, Oyun-Erdene.N, Yanjindulam D, Sodbileg.Sh* **P1-11: Improvement of the Database Performance of a University Information System** *Munkhtuya.D, Naranchimeg.B, Oyun-Erdene.N, Sodbileg.Sh* **P1-12: Quadrupeds Motion Data Collection Method** *Javkhlan Rentsendorj, Erdenebat Budsuren, Baatarbileg Altangerel, Oyun-erdene Namsrai* **P1-13: Installment for Measuring and Sharing PM 2.5 Air Pollution Concentration Through Social Media** *Unursaikhan Batbayar, Sereeter Lodoysamba, Christa Hasenkopf, Joe Flasher*

Thursday, August 1, 2014

FITAT Business meeting (8:30 ~ 12:00)

Time	Topics
08:30 ~ 09:00	**Registration** **Location: Lobby 1**
09:00 ~ 10:10	**Topic: International Cooperation Discussion** **Location: Room 1** Session Chair: Oyun-Erdene and Hoang Do Thanh Tung
10:10 ~ 10:30	**Coffee Break**
10:30 ~ 12:00	**Topic: International Academic Research Discussion** **Location: Room 1** Session Chair: Wang Ling and Mohamed Ezzeldin A. Bashir

Search by Session

- Interactive Session 1
- Keynote Session 1
- Core Database Technologies and Data Mining Techniques
- Communication and Networking
- Data Mining Applications
- Keynote Session 2
- Interactive Session 2
- Keynote Session 3
- Biomedical Informatics
- PSM Session
- PSM, Text Mining and Natural Language Processing
- Communication and Signal Processing
- Interactive Session 3

Session : Interactive Session 1

- PE malDetector: An Intelligent Malware Detection System Based on Learning Algorithms
  Munkhtuya.D, Usukhbayar.B, Uitumen.J, Sodbileg.Sh, Nyamjav.J
- TCP Based Attack Detection Using WEKA Tool
  Ugtakhbayar.N, Enkhjin.B, Sodbileg.Sh
- A Prototype of Expert System for Rural Medical Centers
  Uranchimeg Tudevdagva, Uyanga Sambuu, Yumchmaa Ayush
- Mongolian Traditional Stamp Recognition
  Gantuya Perenleilhundev, Suvdaa Batsuuri
- Adding Mongolian Language Module to the Asterisk Voicemail
  Jamiyan Sukhbaatar, Nyamjav Jambaljav, Bat-Erdene Batdolgor
- Real-time Hand Gesture Recognition using SVM
  Suvdaa Batsuuri, Chintogtokh Batbold
- OOD Metrics for Cohesion - A Survey
  Batnyam Batttulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar
- Implementation of XMI Parser
  Batnyam Battulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar
- Data Mining Techniques Used to Improve Sport Prediction
  Tsend-AyushSh, Otgonnaran O, OyunErdeneNamsrai
- Performance Improvement of Mining Techniques: Supermarket's Data Analysis
  Tsatsral Amarbayasgalan, Bilguun Jargalsaikhan, Otgonnaran O, Oyun-Erdene Namsrai

PE malDetector: An Intelligent Malware Detection System Based on Learning Algorithms

Munkhtuya.D[1], Usukhbayar.B[2], Uitumen.J[1], Sodbileg.Sh[2], Nyamjav.J[2]
[1]*Information Technology Division, National University of Mongolia, Mongolia*
{munkhtuya, uitumen}@num.edu.mn
[2]*School of Engineering and Applied Sciences, National University of Mongolia, Mongolia*
{usukhbayar, sdblg, nyamjav}@num.edu.mn

Abstract

In this work we present a malware detection system using a static hybrid data mining method. We use a combination of two different kinds of features: executable PE header information and dynamic link library (DLL) call features. PE header information were extracted from windows executable PE table, whereas DLL call features were extracted from windows executable PE table's import description table. These features were related each other with one-to-many association. So our work is closely related to relational data mining. In this work, we build a transformation-based approach to relational data mining and implemented Dynamic Aggregation for Relational Attributes (DARA) algorithm that is capable of mapping one-to-many relationship into one-to-one relationship. Then our transformed features were applied to the well-known single-flat classifiers which can classify malicious and benign executable. Finally we implemented an application which can read any executable and then classify them as malicious or benign using trained the classifiers such as SMO and Decision Tree.

1. Introduction

This work investigates the use of data mining methods for malware (malicious programs) detection and proposed a framework as an alternative to the traditional signature detection methods.

The traditional approaches using signatures to detect malicious programs fail for the new and unknown malwares case, where signatures are not available. We present a data mining framework to detect malicious programs. The malware analysis can roughly be divided into static and dynamic analysis. In the static analysis the code of program is examined without actually running the program while in the dynamic analysis the code of program is executed in a real or virtual environment.

In this paper we proposed a static malware detection system using data mining techniques to automatically extract behavior from malicious and clean programs. We collected, analyzed and processed several thousand malicious and clean programs to find out the best features and build models that can classify a given program into a malware or a clean class. Our research is closely related to information retrieval and classification techniques and borrows a number of ideas from the field.

2. Related Works

A similar work is done by Mohammed M.Masud et al. [2]. They extract binary n-gram, assembly n-gram and DLL function call features. We also extract DLL function calls but it differs from [2], in that we extract PE header information. They used 597 benign and 838 malicious executables while we used 271090 malicious and 10592 executables.

In [3], Muazzam Ahmed Siddiqui also presents malware detection system using data mining technique. He extracts the features n-gram and assembly call features like [2].

In [7], Ivan Firdausi et al. presents behavior based malware detection system which can be analyzed on emulated (sandbox) environment. So it is dynamic analysis. They used the classifiers Naïve Bayes, Decision Tree and SVM and reported final detection rates. Our method is content based and based on dynamic analysis.

In [10], Rayner Alfred and Dimitar Kazakov propose multi-instance data transformation approach using DARA. They also propose it in [11] and their accuracy estimation on transformed data is better than

other published results trained on built-in relational data mining systems such as PROGOL, FOIL, and TILDE.

3. Malware Detection System

Our system consists of 5 main parts: (1) data preparation, (2) feature extraction, (3) data transformation, (4) feature selection and (5) classification. Figure 1 shows the architecture of our malware detection system.

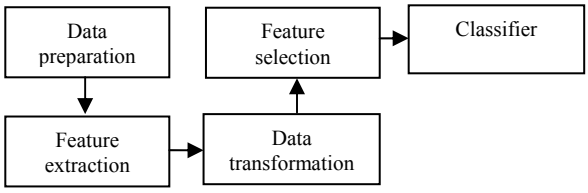


Figure 1. The main process of the system

3.1. Data Preparation

We collected 236756 malicious programs such as backdoor, trojan, virus, worm, exploit, etc and 10592 clean programs. The malicious programs were downloaded from VX heavens web site [4]. There were no duplicate executable. We extracted PE (Portable Executable) header information and DLL call features from these malicious and clean programs and stored in database.

3.2. Feature Extraction

The information about a Win 32 executable file can be easily derived from its PE [9] format without running the program. We developed an application named "PE-Miner" to extract PE header information and DLL call features from the collected executable. This application extracts PE header information, DLLs and their functions list from win32 executable files programmatically. All the PE header information was stored in database. In our database there are totally 236756 malicious executable PE header information, 10592 clean programs PE header information. And 22000 unique DLL call features set consists of both malicious and benign executable.

For simplicity, we only show 4 attributes to present our dataset in Table 1. Where, isBenign is predictive attribute which identifies that an executable file is malicious or clean. The attributes: e_magic, and e_cblp are examples of input attributes. We have totally 112 input attributes. execID is unique ID of executable files.

Table 1. Sample PE header dataset in the database

execID	e_magic	e_cblp	…	isBenign
35	23117	80	⋯	1
52	23117	80	⋯	0
90	23117	144	⋯	0
94	23117	144	⋯	1

Table 1. Sample DLL call feature set

execID	DLLs	Order
80	ADVAPI32.DLL.RegOpenKeyExA	1
80	ADVAPI32.DLL.RegCloseKey	2
80	ADVAPI32.DLL.RegOpenKeyExA	3
80	ADVAPI32.DLL.RegCloseKey	4
80	KERNEL32.DLL.CloseHandle	5
80	KERNEL32.DLL.CreateFileA	6
80	KERNEL32.DLL.CreateProcessA	7
80	KERNEL32.DLL.ExitProcess	8
80	KERNEL32.DLL.FlushFileBuffers	9

In the Table 2, we show one of the execID, which has multiple DLL call features. And, we present "PE-Miner" application program's GUI (Graphic User Interface) in Figure 2.

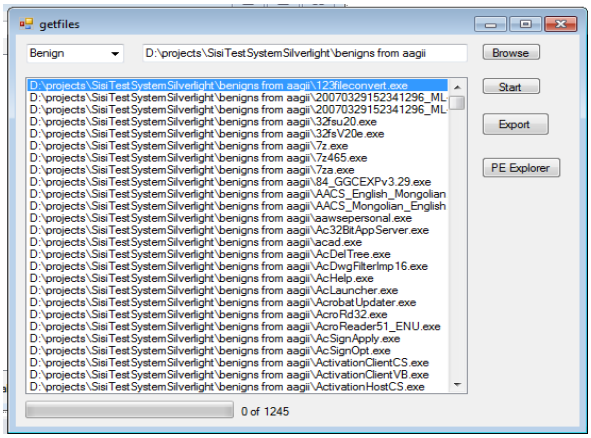


Figure 2. GUI of "PE-Miner" application program: PE header and DLL call features extraction

As we have 2 kinds of features: PE header information and DLL call features. We combine them into one whole hybrid feature set as shown in Figure 3. These features are related each other with the attribute's execID.

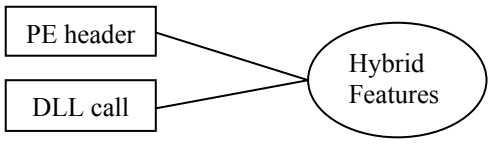


Figure 3. Hybrid feature set

3.3. Data Transformation

Solving the classification problem in relational data mining, traditional methods, for example j48 and its variants, usually require data transformations from datasets stored in multiple tables into a single table. In relational database, a record stored in target table can be associated to one or more records stored in another table due to the one-to-many association constraint [11].

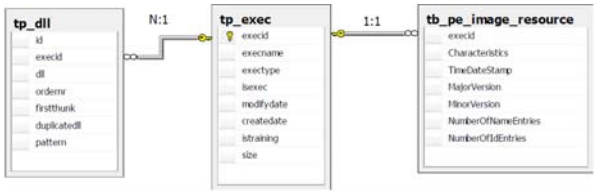


Figure 4. Piece of database diagram

We displayed piece of our database diagram in the Figure 4. For simplicity, we displayed only 3 tables to illustrate our technique. In this case tp_exec is target table and tp_dll and tb_pe_image_resource are non target tables. In this figure, tb_pe_image_resource is one representation of other one-to-one relations.

We joined all the 1:1 relations and created single table tb_all_pe_info. Using traditional data mining tool, multi instance problem is ignored. So, to extract classification rules from relational database with more effectively and efficiently, we need to aggregate these multiple instances. We can treat these multiple instances into bag of terms. There are few ways of transforming these multiple instances into bag of terms.

3.3.1. Vertical Aggregation. Author names and affiliations are to be centered beneath the title and printed in Times 12-point, non-boldface type. Multiple authors may be shown in a two- or three-column format, with their affiliations italicized and centered below their respective names. Include e-mail addresses if possible. Author information should be followed by two 12-point blank lines.

Let Φ denotes the aggregation of multiset values to a categorical or numerical value. A vertical aggregation ΦV is a mapping from multiset values to a categorical or numerical value, λV. It is also known as basic aggregation. Most relational database systems have these aggregation functions for multiset values. For example, aggregation techniques used for categorical value are mode and count. However, these operators (mode, count) can capture only limited discriminative information. On the other hand, relational database's functions like sum, average, count functions can be used to aggregate numerical values. A vertical aggregation can be computed by projecting the column and aggregating the column's multiset values based on a predefined mapping function (sum, count, average, mode), grouped by the target column. For example, suppose we have two relations R1 and R2, with attributes (R1.A1,…,R1.An) and (R2.B1,…,R2.Bn) and R1 has one-to-many (1: n) relationship with R2, then the aggregation Φ^V of attribute Bm, where ($1 \leq m \leq n$), to a categorical or numerical value, λ_V, would be as follows;

$$\Phi^V(\Pi_{R2.Bm}(R1 \; {}_{R1.Ai = R2.Bj} \otimes_{1:n} R2)) \rightarrow \lambda_V \quad (1)$$

Where $1 \leq m, i, j \leq n$, $\Pi_{R2.Bm}$ is the projection of column Bm from the joined relation of R1 and R2 based on the condition R1.Ai = R2.Bj and the λ_V denotes the categorical or numerical value resulted from the aggregation function and $\otimes_{1:n}$ denotes the left join of relations R1 and R2 with one-to-many (1: n) relationship. Notice that when we have one-to-one (1:1) or many-to-one (n:1) relationship, multiset value aggregation will not be required [10].

3.3.2. Pattern-based Feature Aggregation. A common method to aggregate a single categorical attribute with numerous patterns is the selection of a subset of pattern that appears most often or based on the distribution. In this approach, each record is viewed as a vector whose dimensions correspond to patterns stored in the target table in relational domain. The component magnitudes of the vector are the *pf-irf* weights of the patterns which is adapted from [10].

$$pf\text{-}irf = pf(p,r) * ifr(p) \qquad (2)$$

pf-irf is the product of pattern frequency pf(p,r), which refers to the number of times pattern p occurs in the corresponding record r, and the inverse record frequency, as described in (3),

$$ifr(p) = \log \frac{|R|}{rf(p)} \qquad (3)$$

where |R| is the number of records in the table and rf(p) is the number of records in which pattern p occurs at least once. For instance, client X may have three out of ten transactions use cash worth below 150. So, we can say that client X has three occurrence of pattern p (p = 3, |R| = 10), where p is the pattern of making cash transaction with the amount less than 150. The similarity between two records is then

$$sim(r_i, r_j) = \frac{rf * pf}{\|ri\| * \|rj\|} \qquad (4)$$

where r_i and r_j are vectors with *pf-irf* coordinates as described above. As mentioned before, aggregation can be defined as a summarization of the underlying pattern or distribution from which the related objects were sampled. Once we compute the pf-irf weights, then we can compute the distance between each record and cluster them based on their weights. By grouping them into clusters or templates, we are aggregating them based on the underlying pattern or distribution from which the related objects were sampled [10].

3.3.3. Transformation Using Pattern Frequency. Figure 5 shows how a record with bag of multiset values is generalized into a single value. The generalization task is done by computing similarities (4) between records, by first computing the pattern-frequency and inverse-record frequency (2) and then grouping them based on the distance between records. This individual-centered concept, in which a row that belongs to a specific record, illustrates a series of patterns characterize the individualism of each record. As a result, a high-dimensional one-to-many relationship between two tables can be treated as a document [10].

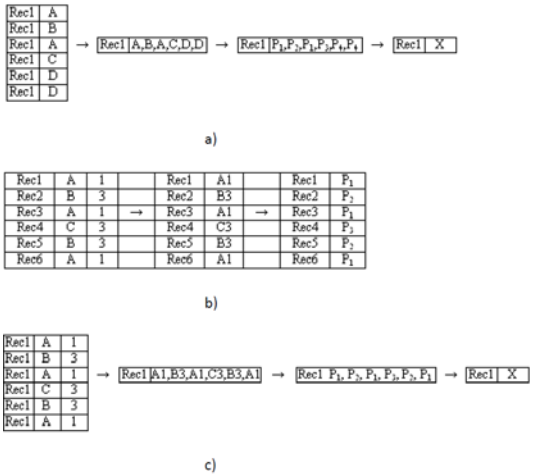


Figure 5. Transformation of frequent pattern using a) Vertical aggregation b) Horizontal aggregation c) Cross aggregation

3.3.4. Implementation. Our database contains a set of relations. One of them is target relation tb_all_pe_info which we have created by joining all the 1:1 relations. The relation tb_all_pe_info has class labels. The other relation tp_dll has no class labels. Each relation has one primary key and several foreign keys. Based on the schema shown in the Figure 7, we would like to classify executable based on characteristics of their DLL call features patterns. Relation tb_all_pe_info has a one-to-many relationship with relation tp_dll.

We used DARA algorithm to generalize each individual in relation tb_all_pe_info that corresponds to the data stored in relation tp_dll and inserted newly generated features into the relation tb_all_pe_info. The main idea of DARA is to generate a cluster that generalizes the relationship between target and non-target tables, and insert the newly generated features into the target table and build the main classifier [10].

In the figure 6, we displayed how multi instance data transformed into single instance. To implement DARA algorithm we followed following steps in DARA algorithm.

Algorithm DARA
Input: A relational database
Output: a set of rules that distinguish the class label.
Procedure:

Rule set R = empty
Create-Pattern();
Compute-Similarity-And-Transform ()
Update-Target-Table ()

Rule r1 = Find-Rule-Target-Table()
Add r1 to the R.
Rule r2 = Find-Rule-Support-Table ()
Add r2 to the R
Return R

End Procedure

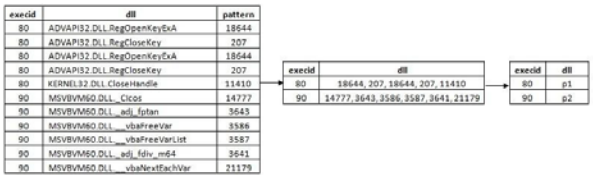


Figure 6. Implementation of Pattern-based transformation using vertical aggregation

3.4. Feature Selection

We used SQL Server Analysis Services (SSAS) to select and reduce features. In general, feature selection works by calculating a score for each attribute, and then selecting only the attributes that have the best scores. You can adjust the threshold for the top scores. Feature selection is always performed before the model is trained, to automatically choose the attributes in a dataset that are most likely to be used in the model. We used Shannon's Entropy to select features and reduce dimensionality of feature space. Shannon's entropy measures the uncertainty of a random variable for a particular outcome. For example, the entropy of a coin toss can be represented as a function of the probability of it coming up heads. SSAS uses the following formula to calculate Shannon's entropy for a random variable X with n outcomes $\{x_i: i=1,\ldots,n\}$

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i) \qquad (5)$$

Where $p(x_i)$ is the probability mass function of outcome x_i.

We have 112 PE header attributes and 3 DLL dataset's attributes. We assumed that DLL dataset's attributes are definitely important. So we only aimed to reduce PE header attributes. We have 88 PE header attributes by using Shannon's Entropy. After reducing PE header feature set, we combined it with DLL call features.

3.5. Classification

We have preprocessed and combined input features set using above methods, it is given as an input to the well-known data mining algorithms for classification. In this section we used 2 algorithms in WEKA [6]: (1) Decision Tree, (2) SVM with cross-validation 10 folds, i.e., the dataset is randomly divided into 10 smaller subsets, where 9 subsets are used for training and 1 subset is used for testing. The process is repeated 10 times for every combination.

To evaluate our system we were interested in several quantities:

1. True Positives (TP): the number of malicious executable examples classified as malicious executable
2. True Negatives (TN): the number of benign programs classified as benign
3. False Positives (FP): the number of benign programs classified as malicious executable
4. False Negatives (FN): the number of malicious executable classified as benign executable

We were interested in the detection rate of the classifiers. In our case this was the percentage of the total malicious programs labeled malicious. We were also interested in the false positive rate. This was the percentage of clean programs which were labeled as malicious, also called false alarms. The Detection Rate is defined as $\frac{TP}{TP+FN}$, False Positive Rate as $\frac{FP}{TN+FP}$, and Overall Accuracy rate as $\frac{TP+TN}{TP+TN+FP+FN}$.

We interested following classifications with feature selection and without feature selection.

Table 3. Result of the classifications with feature selection

	DR (%)	FAR (%)	OAR (%)
Decision Tree	99.8	0.1	99.3
SVM	99.9	0.9	95.8

Table 4. Result of the classification without feature selection

	DR (%)	FAR (%)	OAR (%)
Decision Tree	99.8	0.07	99.5
SVM	99.9	0.9	95.8

Finally, results of the classifications with feature selection and without feature selection are very similar. So we used classification results without feature selection.

4. Implementation

We designed an application which integrated above mentioned approaches. To do this we exported the model trained using j48 with feature selection in WEKA. Then we used WEKA's API (Application Programming Interface) to load and test already trained model. Finally we prepared standalone setup project.

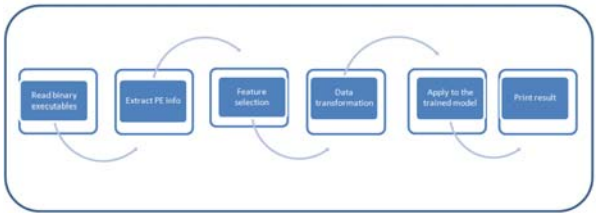
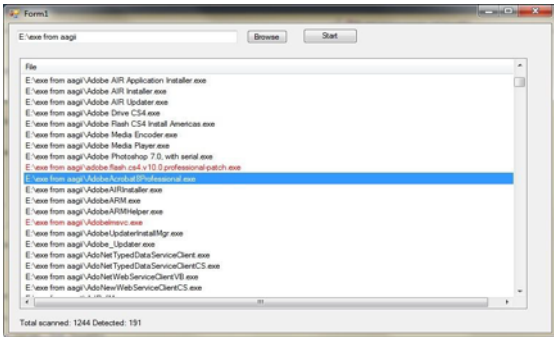


Figure 7. Included steps integration application



5. Conclusion

This work presents a static data mining model to detect malwares. We extracted 2 kinds of features: PE header information and DLL call features from win 32 executable. We have relational database (with multiple tables) and we faced with multi instance dataset problem. The most traditional data mining tools cannot handle relational datasets unless data reduction or data transformation is applied to convert this relational data into single table. So we transformed dataset using patter-based transformation using DARA algorithm.

Then we reduced PE header features' dimension using Shannon's Entropy. Before feature selection we have 112 PE header attributes but by using Shannon's entropy, we have 88 PE header attributes. Then final dataset is given as an input to the well-known classification algorithms such as decision tree, SVM which are included in WEKA. We interested classifications with feature selection and without feature selection. Final results of classifications with feature selection and without feature selection are very similar. Overall accuracy rates are 95.8% - 99.5% and false alarm rates are 0.07%-0.9%. Finally, we integrated all above mentioned approaches in one application to run it in real or virtual environment. We prepared stand-alone setup project.

6. References

[1] Mohammad M.Masud, Latifur Khan, Bhavani Thuraisingham, "A scalable multi-level feature extraction technique to detect malicious executable", 2007.

[2] Mohammad M.Masud, Latifur Khan, Bhavani Thuraisingham, "A hybrid Model to Detect Malicious Executable", *ICC 2007 proceedings*, 1443-1448.

[3] Muazzam Ahmed Siddiqui, "Data Mining Methods for Malware Detection", *PhD dissertation*, University of Central Florida, 2008.

[4] VX heavens web site: http://vx.netlux.org/.

[5] M. Zubair Shafiq, S. Momina Tabish, Fauzan Mirza and Muddassar Farooq, "PE-Miner: Mining Structural Information to Detect Malicious Executable in Real-time".

[6] I. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations". *Morgan Kaufman*, 1999.

[7] Ivan Firdausi, Charles Lim, Alva Ervin, Anto Stariyo Nugroho, "Analysis of Machine Learning Techniques Used in Behavior-based Malware Detection", *2nd international conference on Advances in Computing, Control and Telecommunication Technologies*, 2010, 202-203.

[8] SQL server 2008 R2 books online.

[9] "Microsoft Portable Executable and Common Object File Format Specification" Revision 8.2-September 21, 2010.

[10] Rayner Alfred and Dimitar Kazakov, "Pattern-Based Transformation Approach to Relational Domain Learning Using Dynamic Aggregation for Relational Attributes", *Conference on Data Mining*, 2006.

[11] Rayner Alfred, Dimitar Kazakov, "Aggregating Multiple Instances in Relational Database Using Semi-Supervised Genetic Algorithm-based Clustering Technique", *ADBIS Research Communications*, 2007.

TCP Based Attack Detection Using WEKA Tool

Ugtakhbayar.N[1], Enkhjin.B, Sodbileg.Sh[2]
School of Engineering and Applied Science National University of Mongolia
{ugtakhbayar[1], sdblg[2]}@num.edu.mn

Abstract

In the recent years, TCP based attacks are an important research on computer networking fields. The most dangerous attacks are occurring by DDoS, TCP session attack and exploiting to the operation system and to application processes.

We are introducing by this paper some kinds of method which are learning and comparing collected normal and exploited traffics using artificial intelligent algorithms to detect the attacks. The normal traffic is collected by SNORT tool and the exploited traffic is collected by HOIC tool. The collecting data is de-encapsulating TCP headers, such as sequence number, acknowledge number, window size, control flags and event which is time between neighbor segments. After normalization of the data we apply linear correlation using the Weka tool and also to determine which options impacted to the resultof detecting attacked data.

Keywords*: TCP options; TCP based attack; WEKA; Snort*

1. Introduction

With the development of the information technology number of internet users is increasing fast in the last few years. Within the National University of Mongolia TCP traffic takes xx percent of all traffic. And because of the usage of TCP protocol in most of the stable servers (E-mail, Web etc), TCP attacks occur more than others. The example of this is computers of workers may be used for botnet-based attack.

TCP based attacks such as DoS, session hijacking, exploiting are used the most.

In this experiment we used tools like exploit-db, HOIC, which are used widely worldwide, to attack computers. And then we used WEKA tool for analyzing this traffic. We made feature selection to have the highest rate of detection of attacks and selected Flags, Fragment offset, TTL, Source IP, Destination IP fields from the IP header and Source port, Destination port, Acknowledgement number, Window size fields from the TCP header.

2. Related Work

This type of research has been interested by a lot of people for years. In detection of DDoS attacks, researchers use different types of technique (method), such as the location of an IP address, Inline IPS, Heuristic filtering. In “Botnet-based Distributed Denial of Service (DDOS) Attacks on Web servers: Classification and Art” by EsraaAlomari, they did research on the impact of flood attacks on HTTP. Whereas Sowmyadevi.K described about the impact of gateway devices on DDoS attacks in his paper named “Detect DDoS attack using border gateways and Edge Routers”. Ashwini D. Khairkar included about signature based IDS and detection of web based attacks using ontology in his research “Ontology for Detection of Web Attacks”. Thelow-rate TCP attack is first described by Kuzmanovicand Knightly, who characterize the attackand point out important challenges of detection anddefense.

3. Research Method

In this researchwork, we use a botnet for DDoS attack, Backtrack for exploiting the WEB server, tcpdump and nfdump for collecting the traffic and then WEKA for analyzing attack and normal traffic. We used Pearson correlation method to make less the values of fields and it will increase the rate of detection.

The experiment was made within the School of Applied science and Engineering (SASE) of the National University of Mongolia. The attacking system is Backtrack 5 R3, botnet consists of 10 computers running Windows XP SP2 with the specifications of 1Ghz dual core CPU, 256Mb RAM, 5GB hard,

100Mbps LAN card, the victim is a computer running CentOS 6.2 with the specifications of 2.4GhzCore 2 CPU1GB ram, 40GB hard. And we used web servers of the SASE and the Cisco academy of the National University of Mongolia.

Network traffic was collected by netflow and tcpdump. The filtered traffic by gateway device and watchguard of the National University of Mongolia is collected as a normal traffic. Then we set up a botnet server on experiment devices and while collecting the network traffic we attacked the victim using Backtrack IRC for managing the botnet. The topology used in DDoS attack is shown in figure 1.

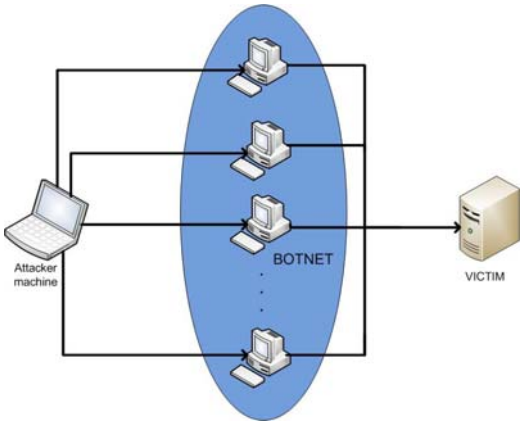


Figure 1. The topology used in DDoS attack

We used HOIC tool for DDoS attacking. Therefore we attacked SASE and Cisco academy web servers while collecting the network traffic by Snort and tcpdump.

4. Solution

At the end of the experiment we collected normal, DDoS attack and exploited traffic. The comparison of the flags in the normal and the attack traffic is shown in the figure 2 and 3. As shown in the figure 2, in normal traffic TCP packets use ACK, PSH, FIN, SYN flags in transmission. But in TCP based attacks flags are used differently. Fast variation of TCP flags means detection by three-way handshake becomes difficult.

In the first experiment we used J-48 algorithm to determine the rate of detection for the TCP flags, TCP ACK, TCP window size fields. Table 1 contains the results.

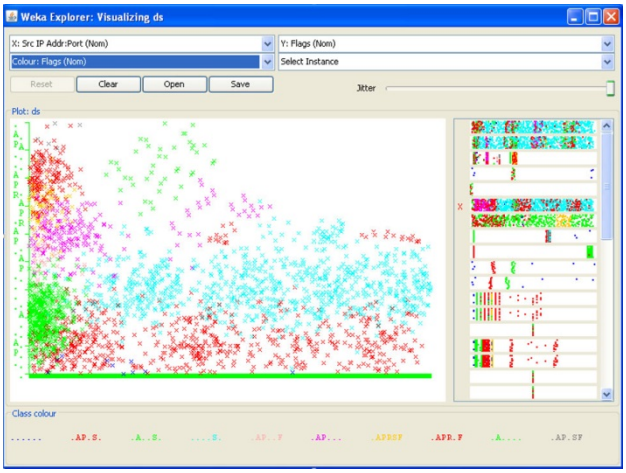


Figure 2. The flag set of attack TCP traffic

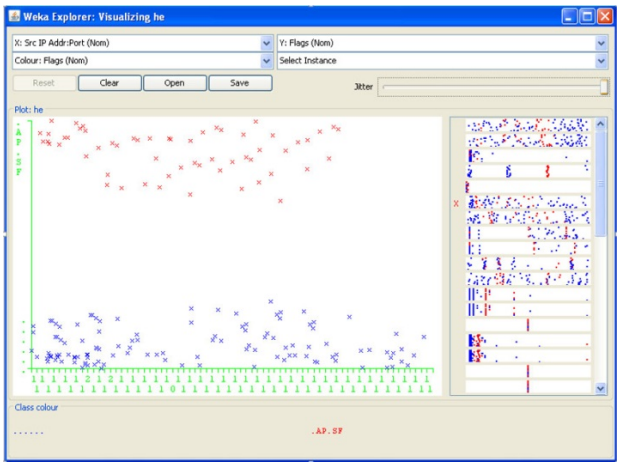


Figure 3. The flag set of normal TCP traffic

Table 1. Detection rate of attack traffic using J-48 algorithm

Feature selection	*TCP_flags*	*TCP_ack*	*TCP_win*
True positive rate(%)	*98%*	*96,8%*	*100%*

Then we used Pearson correlation method (Figure 4) to improve results.

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Figure 4. Pearson correlation

The normal and attack traffic after using Pearson correlation are shown in Table 2 and 3.

Table 2. Attacked traffic

ip_src	ip_dst	time	tcp_sport	tcp_dport	tcp_seq	tcp_ack	tcp_flags	tcp_win	tcp_urp
1	0.043065	-0.3114	0.366409	-0.35599	0.01924	-0.21025	-0.13261	0.219756	0
0.043065	1	0.032226	-0.27542	0.271214	-0.10325	-0.04362	-0.06992	-0.40889	0
-0.3114	0.032226	1	-0.13907	0.099126	-0.38512	-0.2573	0.222455	-0.10153	0
0.366409	-0.27542	-0.13907	1	-0.86014	0.156735	-0.30454	-0.46526	0.512355	0
-0.35599	0.271214	0.099126	-0.86014	1	-0.30785	0.178119	0.2625	-0.38671	0
0.01924	-0.10325	-0.38512	0.156735	-0.30785	1	0.17432	-0.20078	0.28636	0
-0.21025	-0.04362	-0.2573	-0.30454	0.178119	0.17432	1	0.505322	-0.34804	0
-0.13261	-0.06992	0.222455	-0.46526	0.2625	-0.20078	0.505322	1	-0.66496	0
0.219756	-0.40889	-0.10153	0.512355	-0.38671	0.28636	-0.34804	-0.66496	1	0
0	0	0	0	0	0	0	0	0	1

Table 3. Normal traffic

ip_src	ip_dst	time	tcp_sport	tcp_dport	tcp_seq	tcp_ack	tcp_flags	tcp_win	tcp_urp
1	-0.07121	0.549692	0.223295	-0.10574	0.002207	0.165121	-0.13162	-0.13066	0.061697
-0.07121	1	0.54553	-0.18875	0.311146	0.217049	-0.01575	0.074959	-0.04194	0.069881
0.549692	0.54553	1	0.034116	0.18847	0.261126	0.141376	-0.10384	-0.08742	0.137321
0.223295	-0.18875	0.034116	1	-0.96924	-0.463	0.442719	-0.17518	0.059534	-0.09109
-0.10574	0.311146	0.18847	-0.96924	1	0.511442	-0.4119	0.142961	-0.07498	0.106863
0.002207	0.217049	0.261126	-0.463	0.511442	1	-0.26352	-0.04639	-0.11145	0.039886
0.165121	-0.01575	0.141376	0.442719	-0.4119	-0.26352	1	0.083495	-0.12671	-0.02341
-0.13162	0.074959	-0.10384	-0.17518	0.142961	-0.04639	0.083495	1	-0.31938	0.055782
-0.13066	-0.04194	-0.08742	0.059534	-0.07498	-0.11145	-0.12671	-0.31938	1	-0.05166
0.061697	0.069881	0.137321	-0.09109	0.106863	0.039886	-0.02341	0.055782	-0.05166	1

Table 4. Detection rate of attack traffic after using Pearson correlation

Features	*TCP_seq*	*TCP_ack*	*TCP_flags*	*TCP_win*	*TCP_urp*
True positive rate(%)	*88.3*	*89.6*	*58.4*	*84.4*	*80.5*

Next step is to determine detection rate for TCP sequence number, TCP ACK, TCP flags fields using Fuzzy logic and ANNs algorithms. The results are shown in table 5.

Table 5. Detection rate of attack traffic using Fuzzy logic and ANNs algorithms

Algorithms	*Fuzzy logic*	*ANNs*
Feature selection	*TCP_seq, TCP_ack, TCP_flags*	*TCP_seq, TCP_ack, TCP_flags*
True positive rate(%)	*93.5%*	*96.1%*

5. Conclusion

While this research we collected database of attacked and normal traffic. And with this database we determined the detection rate of attacked traffic using J-48, Fuzzy logic and ANNs algorithms. Further we will do feature selection differently to improve the results. Therefore we will determine the detection rate after combine the normal and the attacked traffic for the test dataset.

6. References

[1] Wanli Ma, Dat Tran, Dharmendra Sharma, "A study on the Feature Selection of Network Traffic for Intrusion Detection Purpose", 2008.

[2] Zhi-jun Wu, Jin Lei, Di Yao, Ming-hua Wang, Sarhan M. Musa, " Chaos-based detection of LDoS attacks", *Elsevier journal*, 86, 2013.

[3] A. Kuzmanovic, E. Knightly, "Low-rate TCP-targeted denial of service attacks", *Proc. ACM SIGCOMM*, August 2003.

[4] Haibin Sun, John C.S. Lui, David K.Y. Yau, "Distributed mechanism in detecting and defending against the low-rate TCP attack", *Elseiver journal, Computer networks 50*, 2006 2312-2330.

[5] Annie De Montigny-Leboeuf, "Network traffic flow analysis IEEE CCECE/CCGEI", *Ottawa*, 2006.

[6] Jiao Wang, Yajian Zhou, Yixian Yang, Xinxin Niu, "Classify the Majority of the Total Bytes on the Internet", *International Symposiums on Information Processing*, 2008.

[7] N.Ugtakhbayar, D.Battulga, Sh.Sodbileg, "Classification of artificial intelligence IDS for smurf attack", *IJAIA*, 2012.

A Prototype of Expert System for Rural Medical Centers

Uranchimeg Tudevdagva[1,3], Uyanga Sambuu[2],Yumchmaa Ayush[3]
[1]*Department of Computer Science, Technische Universitaet, Chemnitz,Chemnitz, Germany*
uranchimeg.tudevdagva@informatik.tu-chemnitz.de
[2]*Department of Information and Computer Science, National University of Mongolia, Ulaanbaatar,Mongolia*
uyanga@seas.num.edu.mn
[3]*Power Engineering School, Mongolian University of Science and Technology, Ulaanbaatar, Mongoliayumchmaa@must.edu.mn*

Abstract

This paper describes the prototype of medical expert system for rural medical centers in Mongolia. The rapid development of Information and Communication Technologies (ICT) opens many opportunities to use artificial intelligence in public services such as education, social insurance, customs, taxation,in particular inhealth. The development and use of medical expert system in rural medical centers are essential. In this paper we define key concepts ofmedical expert system and propose prototype of expert system for rural medical centers.The main contribution of this application is to do research on the solution of specific expert system design with targeted group like young doctors and medical techniques engineers in rural medical centers.

Keywords: *Expert System; Rural Medical Center; Application of Expert System; Medical Technique Engineer*

1. Needsand Requirements

At 1,564,116 square kilometers, Mongolia is one of largest landlocked country with population of around 2.9 million people. Given a disperse population in large territory and with widely varying needs, it can be difficult for the government to effectively deliver services to citizens and organizations, especially in rural areas. The problem can become worse as government and population grows but delivery systems do not change. If service delivery becomes slow and uncertain, the cost of delivering services can rise. Besides, there exist some other major problems such as financial constraints, lack of human resource, limited experiences and capacity, and opportunity of decision making, and non-existence of unified standard and universal policy on Governments' Information Technology etc. [2]. To address these problems, the Government of Mongolia attempts to streamline service delivery and bring greater speed, certainty, and transparency to the process.

Government of Mongolia pays special attention to strengthening thehealth medical centers in rural areas.In this regard, significant activities were implemented, such as developing the inter hospital integrated network, testing e-hospital software package, to translate and apply international disease inventories, and review record and reporting templates [10].

The Department of Health is responsible for health statistics information processing.According to the annual report of year 2011, in total 10147 working computers, 3129 fixed telephone are available in 2927 health organizations (1014 of them connected to the Internet) [5]. Most of computers, especially rural areas a hospital and family hospital fails to meet today's need and requirements.

Health organizations are understaffed or lack of trained staff to handle and configure their computers. According to the annual report of year 2011, in total 7943 doctors, 92 statistician doctors and 92 IT staff in 2927 health organizations [10]. There are 4908 doctors, 52 statistician doctors and 61 IT staff in 604 health organizations.

The statistics show that most of health centers are located in rural areas where existing power supply and ICT infrastructure does not satisfy current needs and requirements. The Government recognizes the importance and needs of ICT based services in health sector. In order to strengthenhealth centers we need to develop and implement IT solutions and information systems, health sector-accepted software, appropriate regulations for the health sector information system, system and information security and integrity, expert

system etc.Therefore, we suggest medical expert systemthat support young medical doctors in their professional development and decision making.

2. Expert System

An expert system is the software which helps to solve different problems in various fields. There are key definitions about expert system: *"…Expert system is an intelligent computer program that uses knowledge and inference procedures to solve problems that are difficult enough to require significant human expertise for their solution. An expert system is a computer system that emulates the decision making of a human expert..."* [4], *"…Expert systems are computer applications which embody some non-algorithmic expertise for solving certain types of problems…"* [6].

The expert system is the one of the basic research directions of the artificial intelligence. Exploration of the software with the database was good and basic background for the development of expert systems. With this exploration capacity of database increased dramatically. In order to develop and use of health information database, to analyse complex problems, and to develop and use intelligent algorithms or solutions in decision making processes.

An expert system versus software with the database is included into architecture not only database, also include knowledge base of experts and database for decision making rules.

Usage of the expert system depends on the implementation field and can be different in application. Various types of expert systems are used for decision-making support to define medical diagnosis [1], [3], [7], [8], [9], [11], [12], [13] and treatments. These two main applications are started implement in medicine early with development of artificial algorithms and high technology.

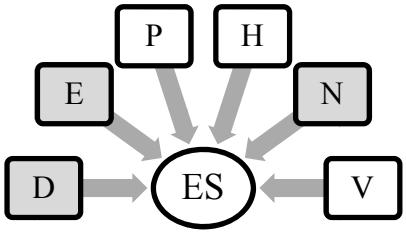


Figure 1. Users of medical expert system

Figure.1shows users of a medical expert systemwhere, ES-expert system, E-engineers, D-doctors, P-patients, H-hospital, N-nurses, V-visitors. This figure can be different depending on the request and targets of the application.

Health organizations in rural areas are understaffed or lack of professionals.Young professionals in most cases prefer to work in urban areas. This circumstance requires the use of ICT in the development of human resource, information and knowledge sharing, and introduction of new technology and best practices. Therefore, we are focusing to the doctors, engineers and nurses. The proposed expert system will support them to optimize decision making by them withoutconsultingwith experts. The decision support expert system supports young doctors and medical engineers to adapt the new environment and to keep motivation to work in rural areas.There are two user groups: doctors and engineers.

Advantages of medical expert systems are:

- Centralized database of the chronically diseases;
- Knowledge database of high level expert's;
- Centralized data bank of patients treatment stories;
- Self improvement opportunity of decision making rules;
- Simplified and improveddecision making processes;
- Time management;
- Accuracy of information and data;
- Knowledge and best practices sharing;
- Support for young doctors and engineersin continuing professional education and research and development.

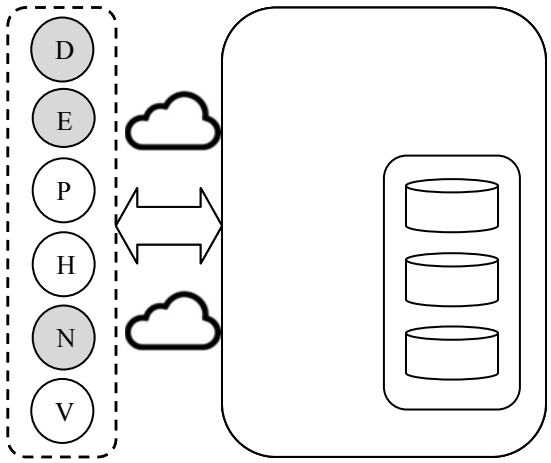


Figure2. Prototype of medical expert system

Limitations are:

- Some major problems such as financial constraints, lack of human resource, infrastructure development;

- Limited experiences and capacity in development and implementation of experts systems;
- Non-existence of unified standard on health information systems.

The proposed prototype of medical expert system is presented in Figure.2 where, Problem – is problem have to solve by doctors or engineers, IA – intelligent algorithm for decision making support, DB – database, KB – knowledge base, Rules – bank of rules for decision making algorithm, EoH – explanation of hints, a storage of all rules which used for offering hints, Hints – main result of experts system, recommendation or hints from expert system to user.

3. Prototype of Medical Expert System

We used a general architecture of expert system that defined early by other researchers. The newest element of prototype architecture is system user interface. Our main challenge of research is to provide user interface not only by the usual visual way, we suggest using cloud environment in connectionof users to the expert system. The Figure.2 showsthe main architecture of a prototype medical expert system for rural medical centers.

The main challenge of our research is to developmedical expert systems in Mongolian universities and to test the proposed prototype. We are planning to develop medical expert system for doctors with special database of specific diseases and expert system for medical engineers with a knowledge base on the complex diagnoses apparatuses.

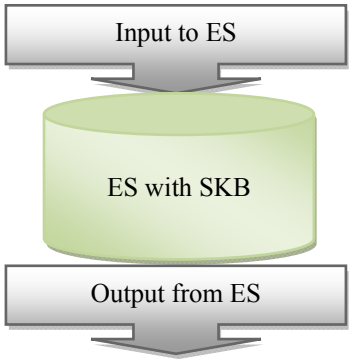


Figure3. Logical concept of ES

The big challenge is the introduction of the user interface by cloud environment. Mongolia has a high probability to use a cloud solution for user interface connection. Figure.2 shows a logical concept of the specific expert system with user oriented version.

This general concept of specific medical expert system can be adapted to user request based on the specific knowledge base (SKB). The pseudocode of above visualized logical concept is below:

```
IF Input to ES is focused to Doctors
      THEN call SKB for Doctors
ELSE
      Call SKB for Medical Engineers
```

The main architecture of the prototype of a medical expert system uses logical concept of expert system presented in Figure. 3.

Our research aims to develop a specific medical expert system with new cloud user interface. The logical concept isa simple and easy to develop and configure for each version: for doctors and medical engineers.

It is also of high priority to reach into collaboration agreement and memorandum of understanding in an area of knowledge base development, information exchange and datasharing to facilitate exchange of information withinhealth organizations and medical experts, and precisely define content, capacity of information to exchange, exchange methods, formats and underpinning regulations.

Key obstaclesare insufficient coordination and coherence over inter-sector information management and organization and lack of IT solutions for data collection, data exchange and data sharing between health organizations and medical experts.

4. Conclusions

The expert system is one of the main research directions of artificial intelligence. Medical expert systems can support medical doctors and engineers of rural medical centers support them in continuing professional education, research and development, information and knowledge sharing, cooperation and collaboration, and decision making.It is essential to build up a modern-standard centralized database of health information, smart or intelligence systems, expert systems, and ensure security and integrity of data.

Future research activities include, but not limited to:

- Literature survey on the existing smart algorithms and selection of smart

algorithm for decision making process for the prototype of a medical expert system;
- Object-oriented analysis and comparison on implementing medical expert systems;
- Architecture design of the specific medical expert system for doctors;
- Development and implementation ofthe specific medical expert system for medical engineers;
- Research and development.

5. References

[1] A.Barr, and E.A. Feigenbaum, "The Handbook of Artificial Intelligence", *William Kaufmann, Inc*., 1, 1981.

[2] Department of Health, Ministry of Health of Mongolia, "Annual report", Ulaanbaatar, 2011.

[3] F. R. Echeverría and C. R. Echeverría, "Application of expert systems in medicine", *Proceedings of the 2006 conference on Artificial Intelligence Research and Development, IOS Press Amsterdam, The Netherlands, The Netherlands,* 2006, 3-4.

[4] J.C.Giarratano and G.Riley, "Expert systems: principles and programming", *PWS-KENT Pub. Co*., 1989, 632 pages.

[5] Information and Communications Technology Authority of Mongolia, Korea IT Industry Promotion Agency, "E-Government Master Plan of Mongolia", April 2005.

[6] D.Merit, "Building Expert Systems in Prolog", *Amzi! Inc*., Online Edition 2000.

[7] M. Mcleish and M. Cecile, "Enhancing medical expert systems with knowledge obtained from statistical data", *Annals of Mathematics and Artificial Intellegence*, 2, 1990, 261-276.

[8] T.Uranchimeg and D.Javzandulam, "An opportunity to use expert system in medicine", *Proceeding of PES, MUST, Ulaanbaatar*, 2006, 25-27.

[9] T.Uranchimeg and N. Munh-Ochir, "Usage of the object oriented method in the decision making process for expert system", *Proceeding of PES, MUST, Ulaanbaatar*, 2009, 15-17.

[10] S.Uyanga, "Baseline Analysis on Current Health Statistics Information System of Mongolia", *The Fifth International Conference on Frontiers of Information Technology Applications and Tools*, 2012, 148-151.

[11] H. Pandza and I. Masić, "Expert systems in medicine", *Pub-Med*, 53 (3 Suppl 3), 1999, 25-7.

[12] A. Yumchmaa, and T. Uranchimeg, "Decision support expert system for medical engineers, problems and solutions", *Proceeding of Health and Young Researchers scientific conference, Ulaanbaatar, Mongolia*, 2009, 19-24.

[13] N. M. A. Zahrani, S. Soomro and A. G. Memon, "Breast Cancer Diagnosis and Treatment of Prophetic Medicine Using Expert System", *Journal of Information & Communication Technology*, 4(2), 2010, 20-26.

Mongolian Traditional Stamp Recognition

Gantuya Perenleilhundev, Suvdaa Batsuuri
School of Engineering and Applied Sciences, National University of Mongolia
gnt1244@gmail.com, suvdaa@num.edu.mn

Abstract

The stamp is one of the crucial information of traditional historical and cultural for nations. In this paper, we purpose to detect and recognize the Mongolian traditional, historical stamps. Thereforewe performed following steps: first, we collected 512 stamp images with 9 classes for training set, second, we implemented the processing algorithms for noise removing, resize and reshape etc. Finally, we proposed a new scale invariant classification algorithm based on KNN(k-nearest neighbor). In the experimental result, our proposed method had shown proper recognition rate.

Keywords: *Stamp Recognition; Image Processing; KNN*

1. Introduction

The cultural behavior, farming, religion, script and fine arts which the ancient Mongolians used to run are expressed as stamp shape.The stamp is significant for defining ancestry, symbol of tribes and is significant for defining the derivation of historical findings.These are consideredto be divided into four things such as pitch-brand, stamps on rocks and tribe stamp.Of the above, the tribe stamp is the most important symbol.We have intended to create database with the image of the tribe stamp and recognize it [1][2]. By researching the tribe stamp, we can create the training database and classify it using a new algorithm.

There are many historical researches that classify the stamps[3]divided into basic and derivative ones according to their founded territories. In this paper we classify them based on their shapes using kNN based algorithm. We have investigated the total of 931 stamps using their basic images.

2. Previous Work

In general there is no specific research that Mongolian stamp recognition using any algorithm from stamp image. Therefore wereviewed the following research works that some sign image recognition works.The KNN algorithm is a popular method for pattern recognition, it has proposed several studies [4], [6], [7].

Also the stamp and sign recognition studies are introduced following researches. Traffic sign detection and recognition [5], Postage stamp recognition using image processing [8], Photo time-stamp detection and recognition [9] and Text detection in images based on unsupervised classification of edge-based features [10].

3. Method

In this paper, we propose a new algorithm based on KNN for stamp recognition. Our system works as follows is shown in Figure 1.

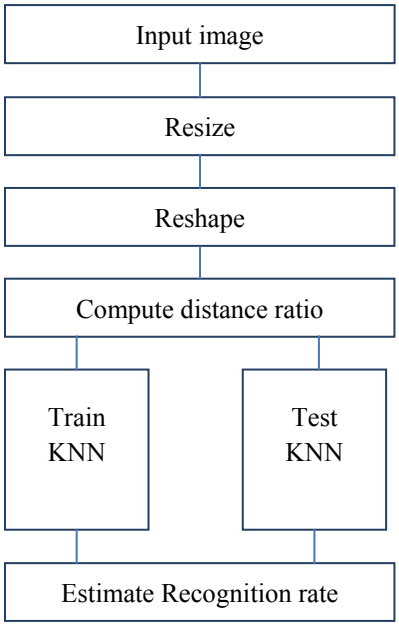


Figure 1. Stamp recognition system

3.1. Introduction to the K-Nearest Neighbor (KNN) algorithm

K-Nearest Neighbor algorithm (KNN) is a method for classifying objects based on the closest training examples in the feature space.[5] KNN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.[6] The KNN algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors (k is a positive integer, typically small)[7]. If k = 1, then the object is simply assigned to the class of its nearest neighbor [8].

3.2. Proposed Method

The proposed method is scale invariant method based on KNN. The original KNN algorithm uses Euclidean distance for comparing the test image and training image from training database. The Euclidean distance is rotation and translation invariant. But it is not scale invariant. Therefore some case, the size of stamp is changed, it fails. Then we propose scale invariant KNN algorithm using distance ratio method as following equation (1).

$$R_{i+1}=d(x_{i+1}, x_{i+2})/d(x_i, x_{i+1}) \qquad (1)$$

Where $d(x_{i+1}, x_{i+2})$ is Euclidean distance between $x+1^{th}$ feature and $x+2^{th}$ features and R_{i+1} is the ratio of those features. This distance is computed for each followed two features.

4. Experimental Results

4.1. Database

We collected 512 stamps of the 9 classes. The following list of the names of those stamps. The images are shown in Figure 2.

1. Onginstamps
2. Tuurai stamps
3. Tahan stamps
4. Saran stamps
5. Saman stamps
6. Builan stamps
7. Zurhen stamps
8. Gurvaljin stamps
9. Alkhan stamps

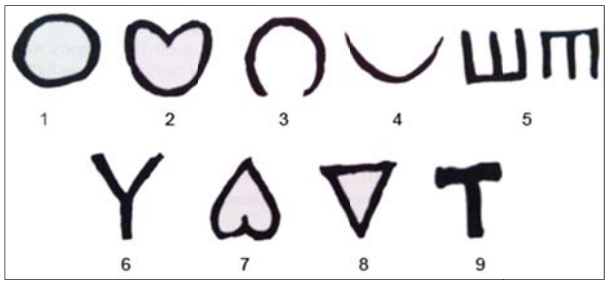


Figure 2. The example stamp images of 9 classes

The experiment processes with steps that have shown in Figure 1. The size of input image is different each other. Therefore we resize the image as size as 20x20. Then we reshape the image matrix (20x20) to vector (400x1). Then in the training phase, we store training images with features distance ratio and in the testing phase, we follow above steps and compute Euclidean distance between the distance ratios. Also we do experiments in k=1 case of KNN.

4.2. Recognition Rate

We test 231 images from training set and 200 images from test set (not in training set). The recognition rate has shown in Table 1.

Table 1. Comparison of recognition rate of two methods

		# of images	Recognition rate (%) of original KNN	Recognition rate (%) of proposed method
K=1	Train error	231	95	100
	Test error	200	90	97

The error of the original KNN was depend on the scale variations. Therefore the new method could improve the recognition rate.

5. Conclusion

In this paper, we purpose to detect and recognize the Mongolian traditional, historical stamps. The proposed method is a scale invariant classification algorithmbased on KNN (k-nearest neighbor).In the framework of this work, we collected 512 stamp images with 9 classes for training set, and tested 231 images from trained images with 100% recognition rate, 200 images are not included in training set with 97% recognition rate.

6. References

[1] Ts. Battulga, “Монголын руни бичгийн бага дурсгалууд”, 2005.

[2] Kh.Perlee, “Бүтээлийн чуулган ”, 2012.

[3] S.Dulam, “Билэгдэл зүй”, 2010.

[4] Hossein Ebrahimpour and Abbas Kouzani, “Face Recognition using bagging KNN”, *Int’l Conf. on CVPR, IEEE*, 1996, 209-216.

[5] Hasan Fleyeh, “Traffic and Road Sign Recognition”, *doctoral thesis, Napier University*, 2008 .

[6] Patrik K, Martina Z, Robert H, Roman J, Miroslav B and Jan H, “A Novel Approach to Face Recognition using Image Segmentation Based on SPCA-KNN Method”, *University of Žilina, Univerzitná 1, 010 26 Žilina, Slovakia*, 2013.

[7] Mohammad K. H., Abu A. S. R., Rajibul A., and Dr.William P., “Automatic Face Recognition System using P-tree and K-Nearest NeighborClassifier”, 2004.

[8] Guancong LiJune, “Postage stamp recognition usingimage processing”, *Faculty of Engineering and Sustainable Development*, 2011.

[9] Xiang-rong Chen and HongJiang Zhang, “Photo time-stamp detection and recognition”, *7th International Conference on Document Analysis and Recognition (ICDAR 2003)*, 2, 2003, 319-322.

[10] Chungmei Liu, Chunheng Wang and Ruwei Dai, “Text detection in images based on unsupervised classification of edge-based features”, *Document Analysis and Recognition, Seoul, Korea*, August 2005, 610–614.

Adding Mongolian Language Module to the Asterisk Voicemail

Jamiyan Sukhbaatar*, Nyamjav Jambaljav*, Bat-Erdene Batdolgor
Department of Electronics and Communication Engineering, School of Applied Sciences and Engineering, National University of Mongolia
{jamiyan,nyamjav}@num.edu.mn

Abstract

Asterisk supports many languages, however, does not support Mongolian. In this paper, we proposed to add the Mongolian language module to the Asterisk voicemail. There are two choices to add Mongolian language. First one is to replace the sound files of Mongolian and the latter one is to add the Mongolian language module. The first choice requires too much resources to implement because of the difference between sentence structures of Mongolian and languages, which supported by Asterisk. Hence, we added the Mongolian language module to the Asterisk voicemail.

Keywords: *Asteris; Voicemai; Mongolian Language*

1. Introduction

Asterisk [1] supports many languages [2], however, does not support Mongolian. The main language is English. The Asterisk is used as an internal communications system at The National University of Mongolia (NUM) [3]. We cannot use voicemail at the NUM because of language problem.

Asterisk based communication systems offer a rich and flexible feature set. Asterisk offers both classical PBX functionality and advanced features, and interoperates with traditional standards-based telephone systems and Voice over IP systems. The list of the features available in Asterisk can be found in [4]. Voicemail is one of the most popular features of Asterisk. The internal communications system of NUM does not use Voicemail of Asterisk due to language problem. The main problem is that sentence structure of Mongolian is different from English and other languages which supported by an asterisk.

The aim of this study is to add the Mongolian language module to the Asterisk voicemail.

2. Asterisk

The asterisk is a complete PBX in software written in C programming language and it runs on many operating systems, including Linux, BSD, Mac and recently even Windows (Windows is however not recommended for this purpose since near real time operation is required). It is an open source toolkit for telephone application and feature-rich callprocessing server. An asterisk can be standalone system or used with previously existing PBX or VoIP implementation. It can be used to manage internal VoIP calls within Local IP Network or in conjunction with specially designed hardware to interface with PSTN networks. Asterisks open source nature causes that it is constantly improving and developing. Currently, Asterisks community counts thousands of participants including programmers, telecommunications professionals, networking professionals and information technology professionals. Such a great support result in a rich set of features, applications and support for all technologies offered by Asterisk platform. Asterisk utilizes many protocols including H.323, SIP and IAX and can interoperate with almost all standard-based telephony equipment using relatively inexpensive hardware. At the same time, almost all leading telephone hardware manufacturers take Asterisk into account and guarantee compatibility with this software.

Asterisk needs no additional hardware for Voice-over-IP, although it does expect a non-standard driver that implements dummy hardware as a non-portable timing mechanism (for certain applications such as conferencing). A single (or multiple) VOIP provider(s) can be used for outgoing and/or incoming calls.

For interconnection with digital and analog telephone equipment, Asterisk supports a number of hardware devices, most notably all of the hardware manufactured by Asterisk's sponsor, Digium. Digium has single and quad span T1 and E1 interfaces for

interconnection to PRI lines and channel banks. In addition, single to quad port analog FXO and FXS cards are available and are popular for small installations. Other vendors' cards can be used for BRI (ISDN2) or quad- and Octo- port BRI based upon CAPI compatible cards or HFC chipset cards.

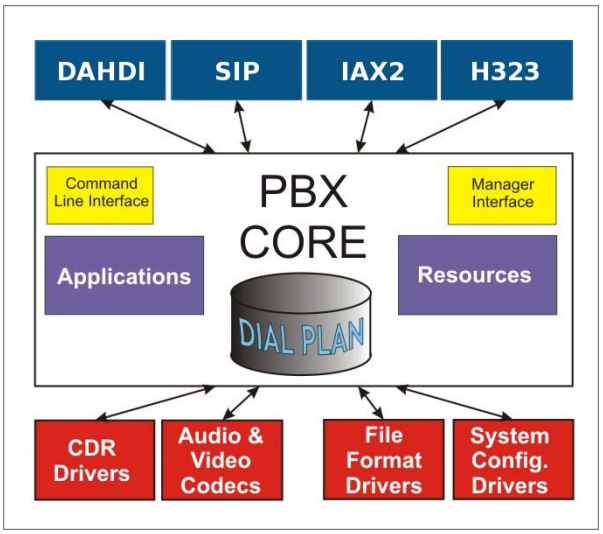


Figure 1. Asterisk architecture

For interconnection with the cellular network (GSM or CDMA), Asterisk can use the Celliax channel driver or chain mobile that is in the trunk now and there is also an unofficial backported version. Lastly, standalone devices are available to do a wide range of tasks including providing fxo and fxs ports that simply plug into the LAN and register to Asterisk as an available device.

3. Voicemail

One of the most popular features of any modern telephone system is voicemail. Asterisk has a reasonably flexible voicemail system named Comedian Mail. Some of the features of Asterisk's voicemail system include:

- Unlimited password-protected voicemail boxes, each containing mailbox folders for organizing voicemail
- Different greetings for busy and unavailable states
- Default and custom greetings
- The ability to associate phones with more than one mailbox and mailboxes with more than one phone
- Email notification of voicemail, with the voicemail optionally attached as a sound file
- Voicemail forwarding and broadcasts
- Message-waiting indicator (flashing light or stuttered dialtone) on many types of phones
- Company directory of employees, based on voicemail boxes

Voice messaging includes several core components. The message collection process is activated when a caller is unable to reach a system user. The message collection application receives data from the phone system indicating which subscriber the caller was attempting to reach. The application plays a greeting, then records the message. The greeting may be a standard system greeting or a custom outgoing message recorded by the subscriber.

Once the message has been recorded, the notification component of the voicemail system takes over and lets the subscriber know that a new message is available. This is handled in different ways depending on the type of phone system with which the voice messaging platform is integrated. In most cases, the voicemail system will send a command to the upstream system (PBX, mobile switching platform, etc.), telling it to turn on the message waiting indicator (MWI) for the subscriber's phone. The notification system may also send an email which may include an audio file attachment of the message.

When the subscriber receives the notification they will access the message using one of several methods. Legacy voice messaging systems require the subscribe to call in an application, authenticate using their extension number and password, and listen to their store messages sequentially. More modern systems allow the subscriber to review their messages on their desktop or mobile phone directly using "visual voicemail." If the message was delivered in an email, the subscriber can listen using their computer as well.
Once the message(s) have been reviewed, the messaging system sends another command to the upstream phone system, instructing it to turn off the message waiting indicator and/or decrease the message count.

Asterisk voicemail mailboxes are defined in /etc/asterisk/voicemail.conf file. The mailboxes might look like the following examples:

```
[exam]
100=>4321,Jamiyan,jamiyan@num.edu.mn
101=>123,Bat-Erdene,baterdene@gmail.com
102=>2222,Dorj,dorj@num.edu.mn
```

There are two primary dialplan [1] applications that are provided by the *app_voicemai.so* module in Asterisk. The first, simply named VoiceMail(), does exactly what you would expect it to, which is to record a message in a mailbox. The second one,

VoiceMailMain( ), allows a caller to log into a mailbox to retrieve messages.

Asterisk dialplan is written in /etc/asterisk/extensions.conf file. In following example, if Jamiyan is busy (on another call) or does not answer the phone for 15 seconds, the caller will be sent to his voicemail, where he will hear Jamiyan's busy or unavailable message:

```
exten => 100,1,Dial(${JAMIYAN},15)
exten => 100,2,VoiceMail(${EXTEN}@exam,u)
exten => 100,102,VoiceMail(${EXTEN}@exam,b)
```

Users can retrieve their voicemail messages, change their voicemail options, and record their voicemail greetings using the VoiceMailMain() application. To allow users to dial 555 to check their voicemail or modify their voicemail options, the following lines should be added to the dialplan:

```
[Incoming]
exten=>555,1,VoiceMailMain()
```

If users dial to 555, it will play a prompt asking the caller to provide her mailbox number and password.

4. Implementation and Results

Voicemail system uses Interactive Voice Response (IVR) [5] system to interact with users. The IVR constructs the sentences using sound files which consist single or multiple words and digits. Figure 2 shows an example of the IVR, when the client received 14 new messages and 7 old messages.

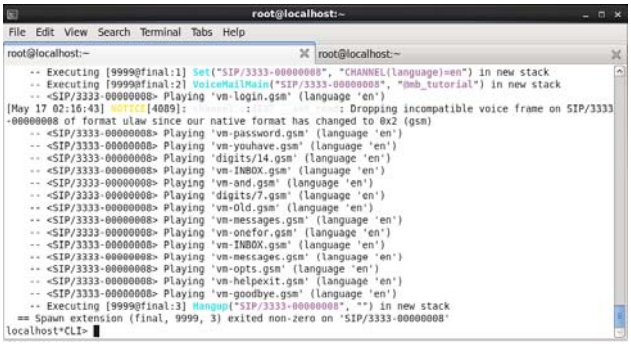


Figure 2. Log of VoiceMail() command

In figure 2, IVR said that "You have 7 new messages and 3 old messages" and to do that eight different sound files have played. This is the main problem to use the voicemail system by Mongolian because of the difference between sentence structure of Mongolian and English. The main difference between Mongolian and English is difference of object and subject of sentence, transaction words and difference of singular and plural. For example, the English sentence structure is "subject + verb + object", whereas Mongolian sentence structure is "subject + object + verb". We wrote Mongolian language module to the Voicemail since, we cannot translate the words which are used to create IVR.

We added Mongolian language functions into the vm_intro(), vm_instructions(), get_folder(), get_play_folder_name(), vm_browse_messages()and vm_execmain() of app_voicemail.c file. Also, functions added into the ast_say_number_full(),which used to say numbers and ast_say_date_with_format() of say.c file. The schematic of voicemail.conf file, Mongolian sound files and Mongolian language module is illustrated in Figure 3.

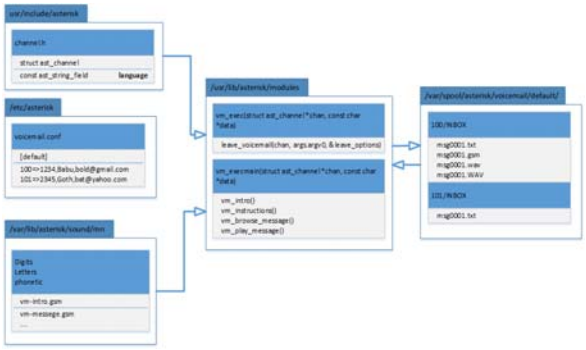


Figure 3. Schematic of VoiceMail module

Following algorithm shows vm_browse_messages() function:

```
static int vm_browse_messages_mn(struct ast_channel *chan,
struct vm_state *vms, struct ast_vm_user *vmu)
{
    int cmd = 0;

    if (vms->lastmsg > -1) {
        cmd = play_message(chan, vmu, vms);
        } else
            {
                cmd = ast_play_and_wait(chan, "vm-tand");
                    if (!cmd)
                        snprintf(vms->fn, sizeof(vms->fn), "vm-%s",
                        vms->curbox);

                        cmd = ast_play_and_wait(chan, vms->fn);
                    if (!cmd)
                        cmd = ast_play_and_wait(chan, "vm-
                        message");
                    if (!cmd)
                        cmd = ast_play_and_wait(chan, "vm-
                        baihgui");
                    if (!cmd)
                        cmd = ast_play_and_wait(chan, "vm-baina");
    }
    return cmd;
}
```

Following dialplan enables Mongolian language module:

```
exten=>100,1,Set(CHANNEL(language)=mn)
exten=>100,2,VoiceMailMain()
exten=>100,3,Hangup()
```

The result of the VoiceMailMain() command illustrated in figure 4.

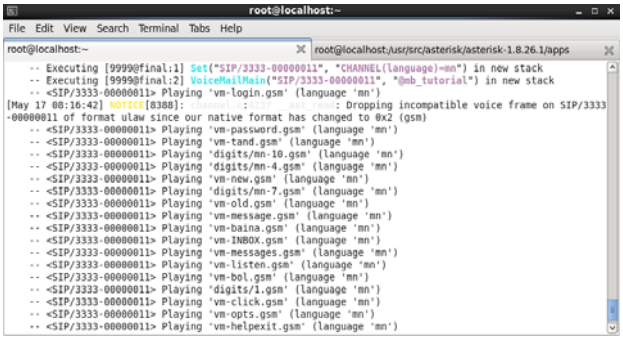


Figure 4. Log of VoiceMailMain() command

5. Conclusion

In this paper, we completed the Mongolian language addition to the Asterisk voicemail. A total of 224 sound files which include words, sentences and 101 digits were used in the Mongolian language module. As a result, we have had the capability to use all possibilities of Asterisk voicemail without any restriction.

6. References

[1] Leif Madsen, Jim Van Meggelen, and Russel Bryant, “Asterisk: The Definitive Guide”, *Third Edition, O’Really*, 2011.

[2] http://www.voip-info.org/wiki/view/Asterisk+sounds.

[3] Jamiyan Sukhbaatar, Batpurev Mongol, “Launching VoIP to the Internal Communications System at the NUM”, *The 5th International Conference FITAT 2012.*

[4] http://www.asterisk.org/get-started/features.

[5] Travis Russell, “Session Initiation Protocol (SIP)” *Controlling Convergent Networking, McGraw-Hill communications*, 2008.

Real-time Hand Gesture Recognition using SVM

Suvdaa Batsuuri, Chintogtokh Batbold
School of Applied Sciences and Engineering, National University of Mongolia, Ulaanbaatar, Mongolia

Abstract

The detection of hands and recognition of hand gestures from both still and moving images is still a matter of research. It has various applications in controlling devices without the use of dedicated physical controllers. Here, we present a method to detect hand gestures from a real time video feed from a normal computer web camera using a combination of computer vision and machine learning techniques.

Keywords: *Hand Gesture Recognition; Computer Vision; Support Vector Machine*

1. Introduction

The successful recognition of hand gestures is a research topic that is still not fully solved. Problems inherent in gesture recognition using two-dimensional video sources include hand direction, camera orientation, obstruction of the hand area by other sources, and source image quality, among many others. Successful recognition presents many opportunities for a wide variety of user applications, such as wireless control of electronic devices without dedicated physical controllers.

Here we present a method of hand gesture recognition from a normal computer webcam using methods of computer vision and support vector machines. We primarily use the OpenCV open source computer vision library [4] in the project..

2. Related Works

Current research of hand gesture recognition involves the following techniques: hand geometry, recognition using Haar-like features, a variety of machine learning techniques, and the use of colored gloved. These techniques are most often used with supplementary techniques such as skin color detection. The commercial release of three-dimensional imaging devices such as the Microsoft Kinect is also opening up the opportunity to use a new spatial dimension in gesture recognition, namely Depth based recognition[11].

All these research projects have had varying amounts of success[9]

3. Method of Detection

3.1. Computer Vision Techniques

The general first step in any method of recognizing human hands involves the segmentation of the image using skin color detection. Naturally, this step has many flaws, including sensitivity to lighting conditions, the false detection of other skin-colored objects, and the variety of human skin colors. Due to the vitality of being independent of lighting conditions, namely brightness, it is useful to convert the image from the RGB color space into a color space that separates luma (brightness) from chrominance (the actual color). The range of skin color in the color space must also be as tight as possible. In any case, most skin color detection research yields a positive rate of 95% and a false positive rate of 15-30%[1]

Most studies indicate a preference in using the YCbCr (where Y is luma and Cb and Cr the blue- and red- difference chroma components respectively) and HSV (hue, saturation, value) color spaces. Images in RGB can be converted to the above two color spaces using a linear formula.

Research yields the following color ranges for human skin color.

- Chai, Ngan [6]: $Cr \leq 173$ AND $Cr \geq 133$ AND $Cb \leq 172$ AND $Cb \geq 77$
- Kukharev, Nowosielski [7]: $Cr \leq 180$ AND $Cr \geq 135$ AND $Cb \leq 135$ AND $Cb \geq 85$ AND $Y \geq 80$
- Oliveira, Conci [9]: $H \leq 50$ AND $S \leq 0.68$ AND $S \geq 0.23$

In addition, bin Abdul Rahman, Wei, and See have proposed a formula using a combination of ranges in the RGB, YCbCr and HSV color spaces, ultimately performing a logical AND on the three components [2]:

- RGB: {R ≥ 95 AND G ≥ 40 AND B ≥ 20} OR {max{R,G,B} − min{R,G,B} > 15) AND (|R−G|>15) AND (R>G) AND (R>B)}
- HSV: G [25,230]
- YCbCr: Cr ≤ 1.5862 × Cb + 20 AND Cr ≥ 0.3448 × Cb + 76.2069 AND Cr ≥ -4.5652 × Cb + 234.5652 AND Cr ≤ -1.15 × Cb + 301.75 AND Cr ≤ -2.2857 × Cb + 432.85

During our testing phase, Chai and Ngan's method proved to be most accurate.

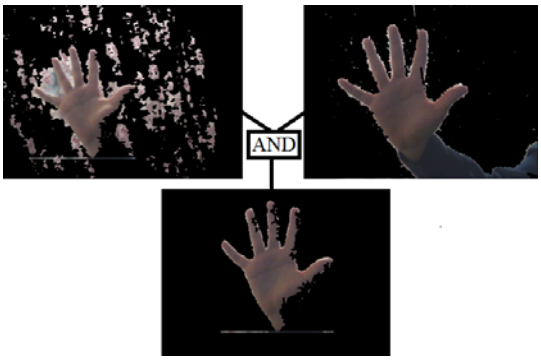
After segmenting the image using skin color, the resulting image contains a large amount of noise. This has been remedied using morphological closing.

3.2. Background segmentation

While skin color detection is a good first step in getting the hand area from the background, it utterly fails when an object of the same color as skin appears in the background. This is very common especially indoors, where most applications of the hand gesture recognition takes place.

Thus an additional step of isolating the hand from the background is necessary. One method of achieving this would be to capture a number of frames from a "blank" background, that is, the image without the hand in it, creating an average model of the background. This can be accomplished by using the Codebook method, whereby a background model is generated by the creation of a codebook that stores pixel ranges over a period of *n* frames. As with skin color detection, RGB is not suitable for codebook generation, thus the preliminary step of converting to the YCbCr color space is taken.[4]

The mask obtained by the codebook method is then combined with the mask obtained from skin color detection, with the masks being logically ANDed, resulting in the final image before other morphological operations are used:

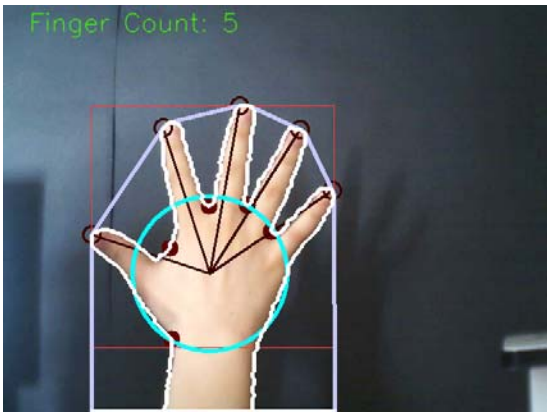


This background segmentation method has an obvious defect in that the camera must remain stationary during use of the program, as well as needing a few seconds before waving the hand in front of the camera to create the background model. However, we deemed it acceptable in the range of our research.

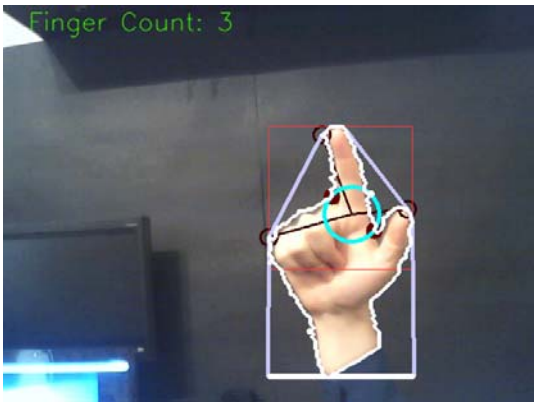
The codebook algorithm we used is based on the implementation in OpenCV.

3.3. Geometric methods

In lieu of using machine learning methods to detect human hand gestures, a rudimentary method of detecting human gestures involves using geometric operations to simply "count" the number of fingers. It involves finding contours from the segmented image (from skin color detection and background segmentation), and drawing a convex hull around the largest contour which the program assumes to be a hand. Then, the convex defects are found, which are assumed to be the points between fingers. Further, a minimum enclosing circle can be drawn from the convex defect points, the center of which is determined to be the hand's center. As a preliminary measure, we created a sample program using this method, with the result below.



As can be seen, this method can correctly determine the number of fingertips, but completely fails todistinguish gestures involving the same amount of fingers but different finger positions, as well as when the gesture shape doesn't conform to the "hand" shape required by this method. As can be seen below, a gesture with thumb and index finger raised is incorrectly counted as having 3 fingers, due to the left part of the hand being slightly bulged outward.



As such, machine learning methods were found to be much more useful and accurate than using a geometric-only model. However, this method still has uses. As it counts contour defects and determines largest contours, parameters can be set so that the program only recognizes a hand as having a range of contour defects, and having an area that falls within a range of largest contour areas.

3.4. Support Vector Machine

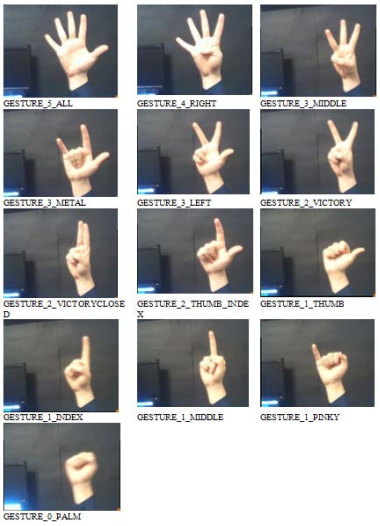
Support Vector machines are a supervised machine learning classifier method used to classify data using a previously learned model. It uses a user-generated model of data to generate an optimal hyperplane to separate the two classifications of data by as wide a margin as possible. Using the kernel method, SVMs can also act as a non linear classifier, essentially transforming the original data into a higher dimension and classifying the generated data linearly in that higher dimension. One of the features of SVMs is that it is possible to calculate the probability of classifying unknown data, unlike other machine learning techniques such as neural networks.[3]

As our data consists of multiple categories, our SVM classifier must be able to distinguish between multiple classifications, known as multiclass SVM. Here, we used the one-against-one technique, which creates a separate SVM for every two categories, as such it is also known as the "pairwise coupling" method. The ultimate decision making step is completed by aggregating the decisions of all the SVMs.[5]

We have used the implementation of Support Vector Machine by OpenCV, itself based on LibSVM.

4. Training and Experimental Results

During the training phase of the project, we first determined the hand gestures we were to use in the project. We settled on 13 gestures, shown below.



Afterward, ~20 second videos in front a blank background were taken of the hand gestures, in both forward and backward orientations (the picture above shows the forward facing gestures), totaling in about 7 minutes of total footage. These videos were put into a program to create two separate SVMs, for forward and backward facing hands.

The SVMs were creates from the videos as follows. The various computer vision techniques outlined in the previous segment were used to segment the hand area from the videos, after which a square was drawn around the hand area. This square was further divided into a 10x10 matrix, essentially resizing the hand into a 10x10px image. The values of the matrix, depending on whether the hand appeared in the matrix element of not, were placed into an array to act as input to the SVM.

After the creation of the SVMs, the test input was used to analyze the success rate of detection. The test input was placed as input into the final gesture

recognition program, and the following data was obtained:

For forward facing:

Gesture type	*Success rate*
GESTURE_5_ALL	100%
GESTURE_4_RIGHT	100%
GESTURE_3_MIDDLE	100%
GESTURE_3_METAL	100%
GESTURE_3_LEFT	100%
GESTURE_2_VICTORY	100%
GESTURE_2_VICTORYCLOSED	100%
GESTURE_2_THUMB_INDEX	100%
GESTURE_1_THUMB	100%
GESTURE_1_INDEX	98.2%
GESTURE_1_MIDDLE	100%
GESTURE_1_PINKY	100%
GESTURE_0_PALM	6.6%

For backward facing:

Gesture type	*Success rate*
GESTURE_5_ALL	100%
GESTURE_4_RIGHT	100%
GESTURE_3_MIDDLE	100%
GESTURE_3_METAL	100%
GESTURE_3_LEFT	100%
GESTURE_2_VICTORY	100%
GESTURE_2_VICTORYCLOSED	97.3%
GESTURE_2_THUMB_INDEX	100%
GESTURE_1_THUMB	15%
GESTURE_1_INDEX	94.4%
GESTURE_1_MIDDLE	100%
GESTURE_1_PINKY	100%
GESTURE_0_PALM	64.3%

As seen, the success rate is very high except for some specific features. Most glaring is the palm, which owes to the fact that the hand as a fist exhibits no contour defects, which the program we developed also checks.

Other gestures with a low success rate have been remedied with an averaging function. The program we developed saves the result of 5 gestures from the previous 5 frames, and determines whether or not the currently detected gesture is valid by checking if it matches at least 3 out of 5 previously detected gestures.

The final phase of our project was to create a simple program to use our gesture recognition results. The below picture shows the program being used to write the words “Hello” in Mongolian. The program accepts the index or middle finger gesture as a cue to draw, and the metal finger gesture as a cue to delete the drawn picture.

10. References

[1] Ahmed, E., M. Crystal, and H. Dunxu, “Skin Detection-a short Tutorial”, *Encyclopedia of Biometrics by Springer-Verlag Berlin Heidelberg*, 2009.

[2] bin Abdul Rahman, Nusirwan Anwar, Kit Chong Wei, and John See, “RGB-H-CbCr Skin Colour Model for Human Face Detection”, *Faculty of Information Technology, Multimedia University*, 2007.

[3] Boswell, Dustin, “Introduction to Support Vector Machines”, *Caltech*, 6 Aug. 2002. Web. 23 Apr. 2014.

[4] Bradski, Gary R., and Adrian Kaehler, “Learning OpenCV: Computer Vision with the OpenCV Library”, *Sebastopol, CA: O'Reilly*, 2008.

[5] Chang, Chih-Chung, and Chih-Jen Lin, “LIBSVM: a library for support vector machines”, *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2.3, 2011, 27.

[6] Chai, Douglas, and King N. Ngan, “Face segmentation using skin-color map in videophone applications”, *Circuits and Systems for Video Technology, IEEE Transactions on 9.4*. 1999, 551-564.

[7] G. Kukharev, A. Novosielski, “Visitor identification - elaborating real time face recognition system”, *Proceedings of 12th Winter School on Computer Graphics (WSCG), Plzen, Czech Republic*, February 2004, 157-164.

[8] Kim, Kyungnam, et al, “Background modeling and subtraction by codebook construction”, *Image Processing*, 2004.

[9] Mittal, Arpit, Andrew Zisserman, and Philip HS Torr, “Hand detection using multiple proposals”, *BMVC*, 2011.

[10] Oliveira, V. A., and A. Conci, “Skin Detection using HSV color space”, *Workshops of Sibgrapi*, 2009.

[11] Ren, Zhou, et al, “Robust hand gesture recognition with kinect sensor”, *Proceedings of the 19th ACM international conference on Multimedia, ACM*, 2011.

[12] Vezhnevets, Vladimir, Vassili Sazonov, and Alla Andreeva, “A survey on pixel-based skin color detection techniques”, *Proc. Graphicon*, 3, 2003.

OOD Metrics for Cohesion - A Survey

BatnyamBattulga, PurevJamai, Naranchimeg Bold, Tamir Chuluunbaatar
Information and Computer Science Department
School of Engineering and Applied Sciences, National University of Mongolia
Ulaanbaatar, Mongolia
{ bbatnyam, purev, naranchimeg, tamir_chuba}@num.edu.mn

Abstract

The designing of highly cohesive classes is an important target in object-oriented design. Class cohesion refers to the relatedness of the class members, and it indicates one important aspect of the class design quality.A meaningful class cohesion metric helps object-oriented software developers detect class design weaknesses and refactor classes accordingly.

With this research, we have considered in the past studies investigated properties of design patterns more precisely and for the evaluation of metrics, we have used not only the static models, but also dynamic models.

The results of this study are the compact summary of the Cohesion Metrics proposed by the scientists and its application possibilities.

Implementation of XMI Parser

Batnyam Battulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar
Information and Computer Science Department
School of Engineering and Applied Sciences, National University of Mongolia
Ulaanbaatar, Mongolia
{ bbatnyam, purev, naranchimeg, tamir_chuba}@num.edu.mn

Abstract

UML tools use XMI format when exchanging information between themselves. Since this format is expressed via XML, it can be expressed in various formats. Although XMI is a common standard, there are certain cases of failure in terms of exchange of data due to its imperfect practical implementation. Hence the purpose of this paper is to implement a parser for most of XMI types that will be able to overcome this shortcoming.

Keywords: *XML Metadata Interchange; XMI Parse; OO Design Metrics*

1. Introduction

XMI standard provides possibilities of effective expression of object modeling using XML language. XMI parser extracts information which is expressed in common modeling language by reading XMI data. Also, basing on the information extracted, it can perform calculation of metrics. Older versions up to the XMI 2.0 define creation of definition of XML document type from model, but newer versions after the XMI 2.0 define creation of XML schema from the model.

Java programming language was mainly used for the performance of the implementation. User interface section was implemented with the help of javafx language. The parcel can also display the metrics calculated in form of html and store them. We believe that this feature will offer ease-of-use and since we are planning to release online version of it, we chose this format.

When performing the user interface, we tried to combine pleasant appearance and simplicity of use. We modeled our HTML content using the Bootstrap framework and performed the user interface window using the FXML.

XMI parser is not only capable of transforming, but it also can perform evaluation. It uses commonly used metrics, criteria, and 15 other different criteria in which some of them consider the work output of other researchers, when performing the evaluation.

The testing version of the software was tested in over 100 class diagrams, 20 use case diagrams, activity diagrams and sequence diagrams as mentioned above. Since the evaluation of diagrams expressed in UML language was a very complicated task, we tested only on class diagrams. There is no permanent normalizations or rules that apply to the value of criteria. Therefore, we determined those values basing upon the results of multiple tests conducted by researchers and also on our own experience. If one wants to change or adjust the values, it would be possible. If we correctly determine the criteria, then we will also be able to evaluate the use case diagram, sequence diagram and activity diagram as well. However, determining the criteria correctly and accurately is a challenging part we face in this stage.

We selected and employed certain languages and technologies considering our further development plan of the software into online accessible version. We should note that since this software was based on the SDmetrics Open Core open source library, it has a full potential of becoming expanded and sophisticated in further. In other words, any desired metrics calculation, rule testing and criteria for evaluation can be updated, added or renewed.

2. XML Metadata Interchange (XMI)

XMI is a standard that expressed the object oriented information in XML format. Commonly, it is used as format for exchange of UML model information. This standard can be applied to all sorts of meta-data in which its meta-models are typically expressed in Meta-Object Family (MOF).
Features of XMI include the following:

- XMI can express the object in XML standard and provides efficient exchange of objects.
- XMI defines how to construct XML schema from the model.
- XMI enables creation of normal XML document as well as such documents of higher level.

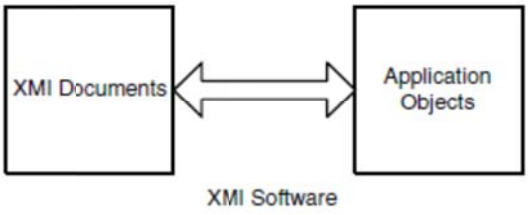


Figure 1. Use of XMI software

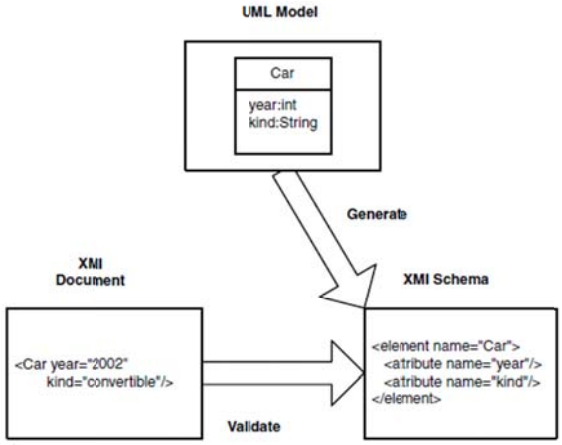


Figure 2. Relation between UML model, XMI document and XMI schema

Meta-Object Family (MOF)

MOF or Meta-Object Family is a standard that defines how to express the meta-data. MOF enables the integration and processing meta-data existing on various levels of virtualization. Also, XMI can be accompanied on MOF defined meta-data existing on various levels.

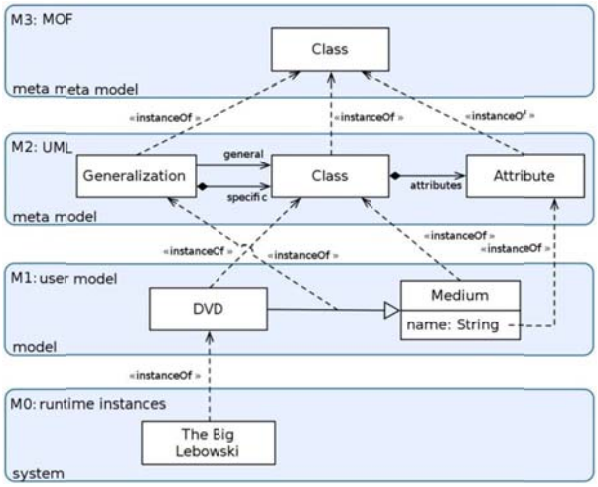


Figure 3. Layer architecture

Version of XMI

XMI is XML namespace compatible. XMI 1.0 and XMI 1.1 versions have been approved on February of 2000. XMI 1.2 and XMI 2.0 version are approved on February of 2001 and XMI 2.0 became compatible of processing XML schema, XML document, and conducting DTD reverse engineering from definition of document type. Subsequently, versions 2.1, 2.1.1, 2.4 and 2.4.1 were released and most recently version 2.4.1 was released on August of 2011. XMI 2.x versions are completely different from XMI 1.x versions. Starting from XMI 2.0, it became possible to conduct mapping of UML models into XML schema. In previous versions, UML models were only able to be gone under mapping into XMI DTD.

XMI 1.0 - XMI DTD, MOF 1.3, UML 1.3, 02/2000
XMI 1.1 - MOF 1.4, UML 1.3
XMI 1.2 - MOF 1.4, UML 1.4 (became schema compatible)
XMI 2.0 - MOF 1.4, UML 1.4 (became schema compatible and syntax was changed) 02/2001
XMI 2.1 - MOF 2.0, UML 2.0, 09/2005
XMI 2.1.1 - MOF 2.0, UML 2.1.1, 12/2007
XMI 2.4 – MOF 2.0, UML 2.4, 03/2011
XMI 2.4.1 – MOF 2.0, UML 2.4.1, 08/2011

3. Implementation

XMI parser was implemented basing on open source library SDMetrics Open Core.

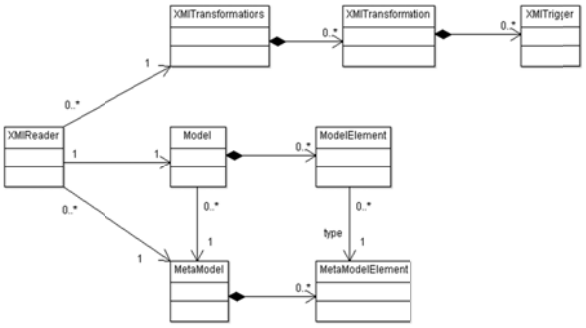


Figure 4. Class diagram for modeling package

Using the SDMetrics Open Core library, each respective element can be accessed and metric calculations associated therewith can be performed.

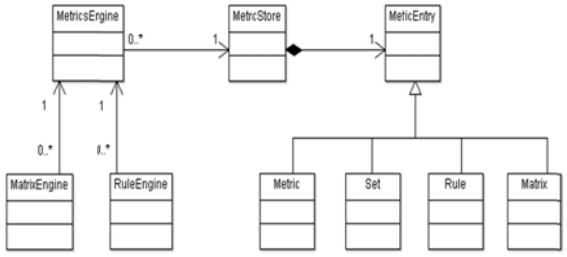


Figure 5. Class diagram for metrics package

Main sections of the software were implemented on Java programming language and user interface was developed using JavaFX, HTML, and Bootstrap CSS technologies.

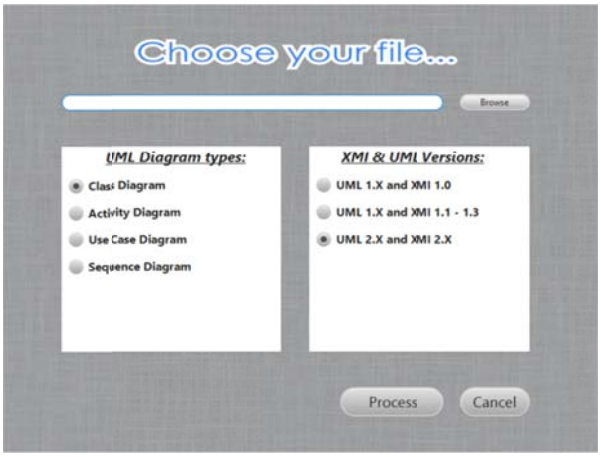


Figure 6. XMI parser GUI

During the implementation process, additional commonly used class diagram metrics such as MOOD, CK, Lorenz & Kidd were also integrated as well as SDMetrics Open Core built-in metrics.

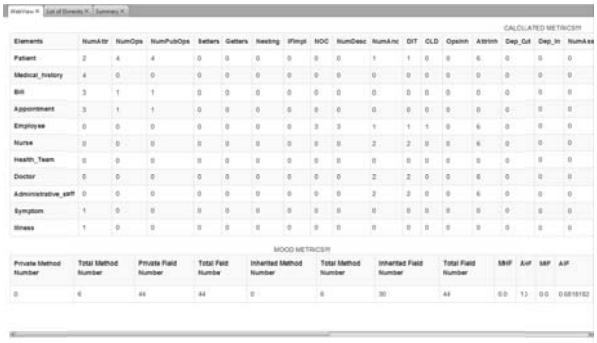


Figure 7. Section showing the listing of metrics

4. Result

Metric calculation was performed on 4 types of diagrams using the implemented XMI parser. Metric calculations were also performed on over 20 use case diagrams, activity diagrams, sequence diagrams and pertinent listing of elements was made. Metric calculations were performed on over 100 class diagrams and relevant listing of elements was made in line with its evaluation.

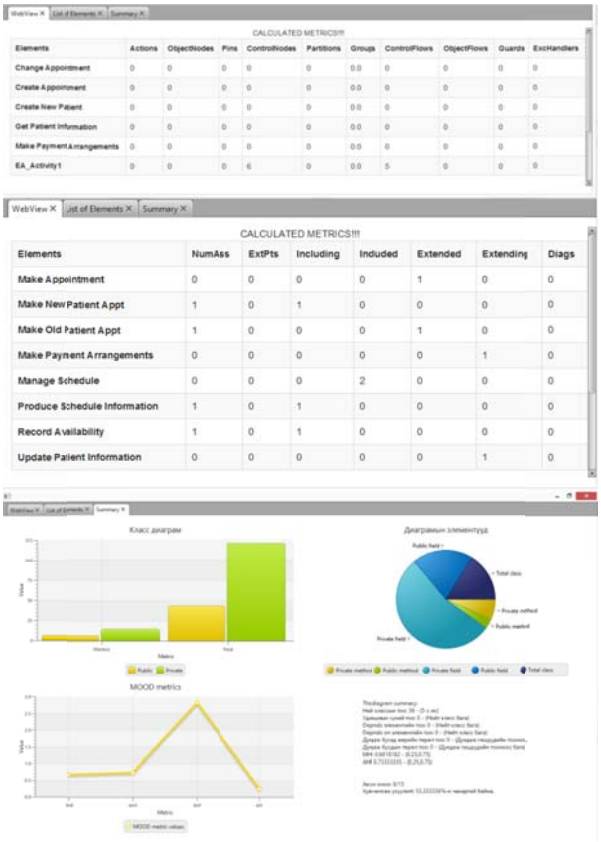


Figure 8. Performing metric calculation of UML diagram

5. Conclusion

Within the scope of this implementation work, a tool which is able to parse UML diagrams expressed as XMI document and to perform Object oriented design metric evaluation on them was developed.The development process was based on SDMetrics Open Core library and used Java programming technology.

Over 100 class diagrams, around 20 use case diagrams, activity diagrams and sequence diagrams were parsed and went under metric calculation. Since performing metric evaluation of all types of UML diagrams was very complicated task, only class diagram evaluation module was developed in this version of the software. If we define criteria for use case diagram, sequence diagram and activity diagram, then the evaluation can be performed without any problems.

In further, metric calculation, rule definition and evaluation criteria can be added to or modified in the software for future development.

6. References

[1] Timothy J, Grose, Gary C. Doney, Stephen A.Brodsky. “Mastering XMI: Java Programming with XMI, XML, and UML”, 2002.

[2] Stephen A.Brodsky, “Object Interchange with XMI”, 2000.

[3] Stevens University of Edinburgh, “XMI and MOF: a mini-tutorial”.

[4] XML Metadata Interchange Specifications (2.0, 2.1, 2.1.1, 2.4, 2.4.1).

[5] XML Metadata Interchange, http://en.wikipedia.org/wiki/XML_Metadata_Interchange, 2014.

[6] XMI official website, [http://www.omg.org/spec/XMI/], 2014.

[7] Thorsten Arendt, Florian Mantz, Gabriele Taentzer, “UML Model Quality Assurance Techniques”, 2009.

[8] *International Journal of Soft Computing and Engineering (IJSCE)*, 2(5), Nov 2012.

[9] Muktamyee Sarker, “An Overview of Object Oriented Design Metrics”, *Master thesis, Department of CS, Umea University, Sweden*, 2005.

[10] Marcela Genero, Martin Piattini and Coral Calero, “Early measures for UML class diagrams”.

[11] Martin Monperrrus, Jean-Marc, Joel Champeau and Brigitte Hoeltzener, “A Model-Driven Measurement Approach”.

Data Mining Techniques Used to Improve Sport Prediction

Tsend-AyushSh, Otgonnaran O, OyunErdeneNamsrai
School of Engineering and Applied Sciences,National University of Mongolia
{tsendayush, otgonnaran, oyunerdene}@num.edu.mn

Abstract

Data mining is a set of actions that summarizes data from various sources, accumulating and processing useful knowledge and information. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

In the era of information technology, data mining is efficiently used not only in the businesses, but also in health and welfare, industrial, social sectors.

This research work focuses on the usage of data mining to predict the result of football match. We developed web application for football fans.This application streams real world football match from predefined sources and streams back using text and audio streaming technology. It also collects FIFA information intelligently from different type of resources and broadcast back to the application users. Application name is “Soccer” and it is now famous for Mongolian football fans. To make our application more interesting we added new module which is intelligently helps for application user to predict result of the incoming match. Several data mining algorithms have been used and presented in this research. We compared prediction rates, and result shows that our module is working well.

Keywords: Sport Prediction; Football Matching; Data Mining; Association Analysis

Performance Improvement of Mining Techniques: Supermarket's Data Analysis

Tsatsral Amarbayasgalan, Bilguun Jargalsaikhan, Otgonnaran O, Oyun-Erdene Namsrai
School of Applied Science and Engineering, National University of Mongolia, Mongolia
{oyun_erdene79, a_tsatsral, ilmaren_9}@yahoo.com

Abstract

Data warehouse collets data from a lot of data sources and store it orderly. That collection of data is used by a decision making systems. The goal of data mining is extract information from large data set and transfer it to comprehensible structure for down the line use. On the other hand, data mining is a method at the using of decision making system and analyzing data, a process of summarizing information when would be effective.

In this paper, we will present that how to model a database to access easily to the large amount of data, preprocessing data using "weka" data mining tool and identifying the association among them. We used real world data of purchases of 2 months in the xxx supermarket.

***Keywords**: Data Warehouse; Data Mining; Data Preprocessing; Association Rule*

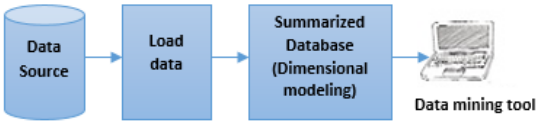
1. Introduction

In the present, accounting programs have made abreast and centrally researches that how to correctly store their existed data, how to analyze them have been making.

Collection of data set is large scale data. Also, the access is extended and it is very difficult to process (store, adjust, search, transfer, analyze and present understandable) via programs which are a traditional data processing and a simple database management system [1][2].

In this paper, we will present a database modeling research for organizing large scale data and experimenting in method of identifying association rule using data mining association algorithms. Certain methods of data mining are used in getting off important information that has possible influence in decision of further business from the large scale data. It is useful for discover interesting relationships hidden in large data sets. For example, it is possible to determine that the customers have bought which product with which one from the accounting of customer's product purchase. Therefore, present supermarket could correctly stage manage own products arranging in row. An another example is that could be get a relationship of which exam is failed students who fail in another which one. Thus, for these students can be pay more attention in that exam that is failed in expected.

Figure 1. Architecture of research



In first section of the background research chapter, we will introduce that how to organize database by a multidimensional model and about its advantages than a relational model, in next sections have been focused on what is the identifying a association rule. In experimental chapter will present that how to modify relational database model of supermarket to the multidimensional model and how to test the method of data mining using "weka" tool.

2. The Background Research

The aim of this chapter is to give the brief insight into the database architecture, data mining association algorithms. It is organized as follows: database dimensional modeling, about data preprocessing, association rule mining, about association rule algorithms in "weka" tool.

2.1. Database Dimensional Modeling

We need to create central repository for extract useful information from it. Data warehouse is integrating data from one or more data sources. So data warehouse include very big data. Database relational model which is traditional approach is suit of database that is related to daily operation. Because the main aim of this model is store not duplicated data and these data related to each other for reduce memory. However it integrates all related tables into

single table when extract useful information from it. This integration step takes more time.

Dimensional model use concept of relational model, but it accelerates performance by decompose data into multi dimension. The main aim of this model is easy and fast access to data but does not focus on duplicated data. In other words, we available to create many cubes instead of tables by dimensional database architecture and extract data from these cubes by using roll up, drill down, dice, slice and pivot operations.

The fact and dimension are concepts that are used to dimensional modeling. The central table is the only table in the schema with multiple joins connecting it to all the other tables. This central table is called the fact table and the other tables are called dimension tables. Numeric values of business operations are stored in fact table and text values are stored in dimension tables. All dimension tables must have one non composite primary key. Fact table include foreign keys that related to primary keys of dimension tables and these foreign keys are joined to create primary key for fact table. But one important thing is fact and dimensional tables have numerical keys that all original keys are replaced by numerical keys. By these numerical keys, fact and dimensional tables are related. Also you must create Time dimension table. It is used to create cube that depends on time.

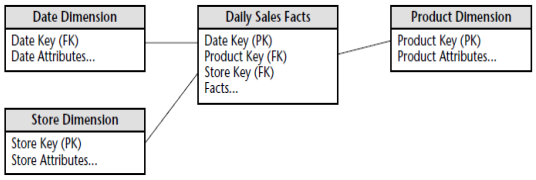


Figure 2. Fact and dimensional tables

Showing Figure 2, Date, Product and Store tables are dimenstional tables and Daily Sales is fact table. All dimensional tables related to fact table and not relate each other. Following table is shown difference between relational model and dimensional model:

Table 1. Difference between relational model and dimensional model

Relational model	**Dimensional model**
Data is stored in RDBMS.	Data is stored in multi- dimensional database.
Unit of data storage is table.	Unit of data storage is cube.
Data is normalized for OLTP.	Data is used to data warehouse and not normalized. Focus on OLAP processing.
Data is temporal (it is modified several time).	Data is not temporal.
Detailed data of operations.	Aggregation result of batch operation for business decision.
Use SQL for data manipulate.	Use MDX for data manipulate.
Simple reports.	User friendly and various reports.

2.2. Data Preprocessing

Data is the collection of objects and objects have attributes. Attribute means feature of object. For example: eye color and height such as features are different from each object. Attribute can be named by another name such as field, variable [3].

Table 2. Data records

Number	Refund	Marital status	Income	Cheat
1	Yes	Single	125000	No
2	No	Married	100000	No
3	No	Single	70000	No
4	Yes	Single	120000	No
5	No	Divorced	95000	Yes
6	No	Married	60000	No
7	Yes	Divorced	220000	No
8	No	Single	85000	Yes
9	No	Married	75000	No
10	No	Single	90000	Yes

Real world data is not ready for used to data mining algorithms. So we prepare compatible data for data mining algorithms by processing raw data. In other words we detect anomaly, missing and duplicated values from data then we erase these problems. We use various strategy and techniques for data preprocessing:

Integration method

By number of attributes into a single attribute to reduce scale of data.

Sampling method

This method select sample subset of whole data. For example: 30% of whole data is related to male and seventy percent is related to female people. In this case male and female people's information is must be balanced. So we do sampling from 70% that is related to female people.

Feature subset selection mehtod

We may reduce scale by remove unimportant attributes or double meaning attributes from data. For example: student's code attribute is not related to compute student's mark.

2.3. Association Rule Mining

In this section, we represented terms of association rule mining. For example:

Table 3. List of purchase

Number	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Egg
3	Bread, Diaper, Beer, Cola
4	Bread, Milk,Diaper, Beer
5	Bread, Milk, Diaper, Cola

For example we can extract following rule from Table 3 {Diaper} → {Beer}
From above rule, can be found strong association between diaper and beer. Because most customer buy diaper with beer [3]. Association rule mining is not used only market basket data, can be used to bioinformatics, hospital diagnostic, web mining and other special science. This paper focus on market basket data and analysis it.

Binary representation

The first step of association rule mining is that represent data into binary representation. Table 4 shows market basket data. In this table each row is a transaction and columns are purchased items. If customer buy an item corresponding column value is 1, else value is 0. In other words, it is more important whether the product has been involved in the purchase rather than how many times the product has been purchased.

Table 4. Binary representation

Num-ber	Bread	Milk	Diaper	Beer	Cola
1	1	1	0	0	0
2	1	0	1	1	0
3	0	1	1	0	1
4	1	1	1	0	0
5	1	1	1	0	1

Terms of association rule:
Item set
Collection of one or more items.
For example: $\{Bread, Diaper, Beer\}$
Association rule
It is represented by $X \rightarrow Y$ expression form. X and Y is set of items.
For example: $\{Bread, Diaper\} \rightarrow \{Beer\}$
Support count(σ)
Frequency of occurrence of an itemset. For example: $\sigma(\{Bread, Milk\}) = 3$ (Table 4 was taken in this example)
Support
Fraction of transactions that contain an itemset.
For example: $s(\{Bread, Milk\}) = 3/5$
Frequent itemset
An itemset whose support is greater than or equal to a minsup threshold (σ(support count)).
Rule evaluation metrics
There are support and confidence metrics in rule evaluation metric. In the below example, shows how to measure $\{Milk, Diaper\} \rightarrow \{Beer\}$ rule?

- Support (s)
- Fraction of transactions that contain an itemset.

$$s = \frac{\sigma(\{Milk, Diaper, Beer\})}{|T|} = \frac{1}{5} = 0.2$$

- Confidence (c)
  Measures how often items in Y appear in transactions that contain X.

$$c = \frac{\sigma(\{Milk, Diaper, Beer\})}{\sigma(\{Milk, Diaper\})} = \frac{1}{2} = 0.5$$

2.4. About Association Rule Algorithms in "weka" Tool

There are many algorithms for association rule. The most famous of these algorithms is Apriori. In this section, Appriori, FilteredAssociator, Predictive Apriori, Tertius algorithms that are used in "weka" described briefly.

Apriori algorithm

The purpose of this algorithm is to find subsets which are common to at least a minimum number C (Confidence Threshold) of the itemsets.Terms related to this algorithm are as follows [6]:
Frequent itemsets: The sets of item which has minimum support and it is denoted by L_i for i^{th} itemset.
Apriori property: Any subset of frequent itemset must be frequent.
Join operation: To find L_k , a set of candidate k-itemsets is generated by joining L_{k-1} with itself.
Join step: Candidate item C_k is generated by joining $L_{k-1}$with itself.
Prune step: Any (k-1)–itemsetthat is not frequent cannot be a subset of a frequent k-itemset.
***Input*:** Database of Transactions $D = \{t1, t2, ..., tn\}$,
Set if Items $I = \{I1, I2, ..., Ik\}$,
Frequent itemset L,
Support,
Confidence.
Output: Association rule satisfying support and confidence.
***Method*:**
1. C_1 Itemsets of size one in I;
2. Determine all large itemsets of size 1, L_1;
3. i = 1;
4. Repeat
5. i = i + 1;

6. C_i = Apriori-Gen(L_{i-1});
7. Apriori-Gen(L_{i-1})
   1. Generate candidates of size i+1 from large itemsets of size i.
   2. Join large itemsets of size i if they agree on i-1.
   3. Prune candidates who have subsets that are not large.
8. Count C_i to determine L_i;
9. until no more large itemsets found;

The best rule from the item set L= {2, 3, 5} are calculated as follows:

Consider the minimum support is 2 and minimum confidence is 70%.

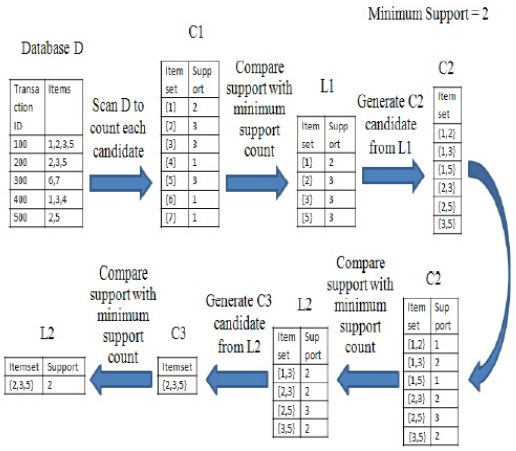


Figure 3. Generation item set & frequent item set

Rule 1: {2, 3} → {5} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2, 3}) = 2/2 = 100%

Rule 2: {2, 5} → {3} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2, 5}) = 2/3 = 67%

Rule 3: {3, 5} → {2} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({3, 5}) = 2/2 = 100%

Rule 4: {2} → {3, 5} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2}) = 2/3 = 67%

Rule 5: {3} → {2, 5} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({3}) = 2/3 = 67%

Rule 6: {5} → {2, 3} Confidence = Support Count of ({2, 3, 5})/ Support Count of ({5}) = 2/3 = 67%

Hence the accepted rules are Rule 1 and Rule 3 as the confidence of these rules is greater than 70%.

Predictive Apriorialgorithm

In predictive Apriori association rule algorithm, support & confidence is combined into a single measure called predictive accuracy. This predictive accuracy is used to generate the Apriori association rule. In Weka, this algorithm generates n best association rule based on n selected by the user.

Tertius algorithm

This algorithm finds the rule according to the confirmation measures (P. A. Flach, N. Lachiche 1999). It uses first order logic representation. It includes various option like class Index, classification, confirmation Threshold, confirmation Values, frequency Threshold, horn Clauses, missing Values, negation, noise Threshold, number Literals, repeat Literals, roc Analysis, values Output etc.

Filtered algorithm

This algorithm finds the rule according to the confirmation measures (P. A. Flach, N. Lachiche 1999). It uses first order logic representation. It includes various option like class Index, classification, confirmation Threshold, confirmation Values, frequency Threshold, horn Clauses, missing Values, negation, noise Threshold, number Literals, repeat Literals, roc Analysis, values Output etc. processed by the filter without changing their structure. Here in this algorithm we can consider the Apriori, Predictive Apriori and Tertius association rule algorithm to get the result.

3. Experiment

There are 2 sections in experiment. The first section shown how to modify database architecture of xxx supermarket into dimensional modeling, next section shown experiment of association rule mining on the data of purchase that was been selling during 2 months in the xxx supermarket.

3.1. Modify Database Architecture of xxx Supermarket by Dimensional Model

The database architecture of xxx supermarket was designed by relational model and **Error! Reference source not found.** shown relational tables in that database. These tables only related to daily transaction, not store other data. In other words, irrelative attributes with transaction is not shown in Figure 1**Error! Reference source not found.**.

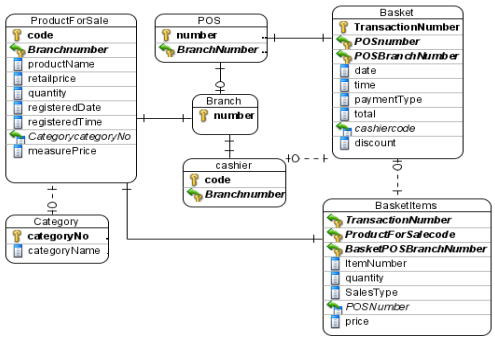


Figure 4. Old database architecture of XXX supermarket /relational model/

Description of tables:

ProductForSale –it stores product code, branch number, product name, retail price, disbursing price, quantity, category which data related to product. It has foreign keys that are related to Branch and Category tables.
POS –it stores about sales points, and has foreign key that is related to Branch table.
Basket –it stores about transaction, and has foreign keys that are related to POS and Cashier tables.
Branch –it stores about branch information of supermarket.
Cashier–it stores about cashier of supermarket and has foreign key that is related to Branch table.
Category –it stores about category of product.
BasketItems –it stores basket information of transaction and has foreign keys that are related to Basket and ProductForSale tables. Figure 5 shown how to transform above relational model into dimensional model.

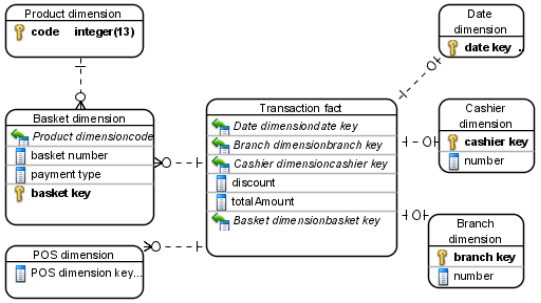


Figure 5. New database architecture of XXX supermarket /Dimensional modeling/

3.2. Find Association Rule from Data Using "weka" Tool

We did our experiment on January and February transaction of xxx supermarket. This data is stored in excel file. Steps illustrating how to convert it into binary representation with the aim to carry out tests:
Step 1: Sold products have been assigned to an appropriate category. There are 43 product categories have been created. In other words, a row designated for the product categories have been inserted whereby a 1 assigned to a sold product and a 0 assigned to the ones not sold.

Bill	drink	sweetne	noodles	prepared fl	prepared f	pastry
1	0	0	0	0	0	0
	0	0	0	0	0	0
2	0	0	0	0	0	0
2	0	0	0	0	0	1
2	0	0	0	0	0	1
	0	0	0	0	0	0

Figure 6. Binary representation of sold products

Figure 6 shows the varying numbers on the rows depending the amount of products sold. For example, 1 piece has been sold from the products assigned with 1 and 3 pieces has been sold from the products assigned with 2.
Step 2: The data has been inserted into the table in the server. From here, every sale can be represented one row by using sql query.
Next all 0 value was replaced with question mark and other values replaced TRUE value. Question mark means that product category did not appear in purchase.
Next, the table created on the data server has been exported into an excel file.
Step 3: the exported excel file has been saved in csv format.
Step 4: open the csv format file with notepad and convert the contents into an arff file's structure. The data preparing process ends here.

```
@attribute beer {TRUE}
@attribute vodka {TRUE}
@attribute biscuit {TRUE}
@attribute sweet {TRUE}
@attribute chocolate {TRUE}
```

Figure 7. 43 attributes have been described in arff file

```
@data
?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
?,?,?,?,?,TRUE,TRUE,?,?,?,TRUE,
?,?,?,?,TRUE,TRUE,?,?,?,?,?,TRU
```

Figure 8. Data in arrf file

A total of 59029 lines and 43 attributes have been created.

Test results:
Apriori Association rule:
/ tests involving all attributes /
1. Minimum support: 0.01
2. Minimum metric <confidence>:0.5
3. Number of cycles performed: 20

Best rules found:
1. Milk = TRUE juice = TRUE sausage = TRUE 888 ==> pastry = TRUE 644 conf:(0.73)
2. PreservedAndCanned = TRUE juice= TRUE driedCurd = TRUE 827 ==> pastry=TRUE 595 conf:(0.72)
3. PreservedAndCanned=TRUE juice=TRUE sausage = TRUE 1072 ==> pastry=TRUE 761 conf:(0.71)
4. egg=TRUE juice=TRUE sausage=TRUE 855 ==> pastry=TRUE 603 conf:(0.71)
5. PreservedAndCanned = TRUE milk = TRUE juice = TRUE 947 ==> pastry = TRUE 667 conf:(0.7)
6. vegetables=TRUE milk=TRUE juice=TRUE 897 ==> pastry=TRUE 629 conf:(0.7)
7. fruit=TRUE juice=TRUE sausage=TRUE 895 ==> pastry=TRUE 623 conf:(0.7)
8. fruit=TRUE milk=TRUE juice=TRUE 867 ==> pastry=TRUE 597 conf:(0.69)

9. vegetables=TRUE juice=TRUE sausage=TRUE 981 ==> pastry=TRUE 674 conf:(0.69)
10. sausage=TRUE driedCurd=TRUE 1304 ==> pastry=TRUE 894 conf:(0.69)

/ test excluding the attributes pastry /

1. Minimum support: 0.01
2. Minimum metric <confidence>:0.5
3. Number of cycles performed: 20

Best rules found:

1. meat=TRUE egg=TRUE 1056 ==> vegetables=TRUE 664 conf:(0.63)
2. fruit=TRUE meat=TRUE 1241 ==> vegetables=TRUE 774 conf:(0.62)
3. chocolate=TRUE fruit=TRUE 1000 ==> juice=TRUE 608 conf:(0.61)
4. PreservedAndCanned=TRUE driedCurd=TRUE 1423 ==>juice=TRUE 827 conf:(0.58)
5. chocolate=TRUE PreservedAndCanned=TRUE 1046 ==> juice=TRUE 602 conf:(0.58)
6. fruit=TRUE egg=TRUE 1303 ==> vegetables=TRUE 747 conf:(0.57)
7. PreservedAndCanned=TRUE fruit=TRUE 1662 ==> juice=TRUE 948 conf:(0.57)
8. sweet=TRUE PreservedAndCanned=TRUE 1038 ==> juice=TRUE 591 conf:(0.57)
9. sausage=TRUE driedCurd=TRUE 1304 ==> juice=TRUE 742 conf:(0.57)
10. fruit=TRUE driedCurd=TRUE 1532 ==> juice=TRUE 865 conf:(0.56)

/ test excluding the attributes of pastry, vegetables /

1. Minimum support: 0.01
2. Minimum metric <confidence>:0.5
3. Number of cycles performed: 20

Best rules found:

1. chocolate=TRUE fruit=TRUE 1000 ==> juice=TRUE 608 conf:(0.61)
2. PreservedAndCanned=TRUE driedCurd=TRUE 1423 ==> juice=TRUE 827 conf:(0.58)
3. chocolate=TRUE PreservedAndCanned=TRUE 1046 ==> juice=TRUE 602 conf:(0.58)
4. PreservedAndCanned=TRUE fruit=TRUE 1662 ==> juice=TRUE 948 conf:(0.57)
5. sweet=TRUE PreservedAndCanned=TRUE 1038 ==> juice=TRUE 591 conf:(0.57)
6. sausage=TRUE driedCurd=TRUE 1304 ==> juice=TRUE 742 conf:(0.57)
7. fruit=TRUE driedCurd=TRUE 1532 ==> juice=TRUE 865 conf:(0.56)
8. egg=TRUE driedCurd=TRUE 1239 ==> juice=TRUE 698 conf:(0.56)
9. fruit=TRUE milk=TRUE 1555 ==> juice=TRUE 867 conf:(0.56)
10. fruit=TRUE sausage=TRUE 1621 ==> juice=TRUE 895 conf:(0.55)

/ test excluding the attributes of pastry, vegetables /

1. Minimum support: 0.01
2. Minimum metric <confidence>:0.5
3. Number of cycles performed: 20

Best rules found:

1. milk=TRUE sausage=TRUE 1622 ==> bread=TRUE 826 conf:(0.51)
2. egg=TRUE driedCurd=TRUE 1239 ==> milk=TRUE 662 conf:(0.53)
3. PreservedAndCanned=TRUE egg=TRUE 1536 ==> sausage=TRUE 772 conf:(0.5)

If we set too low minimum support, result will be false. Let's look at an example where the minimum support count is set to 1. We see that chocolates and potatoes have been sold once each; and when they are sold together, the association is accepted. Therefore, if we consider confidence levels, the result will be 100% when a purchase of chocolate involves potatoes. However, we cannot assume the association to be 100% based on a single purchase.

Our experiment has been set minimum support as 0.01. This means we must discover association of products that have been involved in 0.01% of total 59029 purchases. In the above experiment, association rules have been discovered from products that involved around 590 purchase.

Association rule 1:

The probability of pastry being purchased when milk, juice and meat are purchased is 73%.

milk=TRUE juice=TRUE sausage=TRUE 888 ==> pastry=TRUE 644 conf:(0.73)

Association rule2:

The probability of groceries being purchased when meat and eggs are purchased is 63%.

meat=TRUE egg=TRUE 1056 ==> vegetables=TRUE 664 conf:(0.63)

Association rule3:

The probability of bread being purchased when milk and meat are purchased is 51%.

milk=TRUE sausage=TRUE 1622 ==> bread=TRUE 826 conf:(0.51)

4. Summary

With this study, we have examined what kind of database model need for a large scale data. Also we have reached the conclusion the dimensional modeling is suitable for data that is collected from various sources and relational model is suitable for data of daily operations. Secondly, we have studied the level of association of the products in the daily shopping carts. By doing this, we have gained experience by calculating the actual probabilities and understanding the concepts behind it.

5. References

[1] http://en.wikipedia.org/wiki/Data_mining.

[2] "Big data", http://en.wikipedia.org/wiki/Big_data.

[3] M. S. V. K. Pang-Ning Tan, "Introduction to data mining", *Instock*, 2006, 19-88.

[4] M. J. Swasti Singhal, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering", *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 251, May 2013.

[5] Machine Learning Group at the University of Waikato, "Data Mining Software in Java", http://www.cs.waikato.ac.nz/~ml/weka/index.html.

[6] L. L. Sunita B. Aher, "A Comparative Study of Association Rule Algorithms for Course Recommender System in E-learning", *International Journal of Computer Applications*, 39(1), February 2012.

Session : FITAT Opening and Keynote Session 1

- Understanding EEG Signal for Better Brain-Computer Interface and other Applications

  Goutam Chakraborty

- Fuzzy Set Theory in Information Technology

  Sansanee Auephanwiriyakul

Understanding EEG Signal for Better Brain-Computer Interface and other Applications

Goutam Chakraborty
Iwate Prefectural University
goutam@iwate-pu.ac.jp

Abstract

Brain Computer Interface (BCI) or Brain Machine Interface (BMI)are devices that facilitate communication or interaction by analyzing electrical signals collected by probes set on the scalp or inserted inside the head touching the brain. Electroencephalography (EEG) is recording of electrical potential on the scalp. Analysis of the EEG to interpret the intention of the user, and use that for communication or maneuveringa machine, is the core idea of BCI.

With the advent of low-noise high-sensitivity probes and cheap powerful computers, collection of brain-signals from multiple electrodesand their analysis in real-time,is possible. This is leading to a lotof attention to the research and development of BCI applications. Not only medicalapplications (like epileptic seizures, monitoring anesthesia or brain function etc.), but also applications like moving a wheelchair, or communicating using BCI speller, are getting more accurate and affordable. There are entertainment applications too, for those who can not move their limbs.

The main problem with present BCI application tool is the fact that the number of probes needed is many. A better understanding of the EEG signals, generated during various BCI applications, is the key to reduce the number of required probes, and make their use more common.

In this talk, we will first explain how we could reduce the number of probes by designing it for an individual. In other words, by exploiting individual variation, we can optimize the number of probes without sacrificing the quality. In the second part of the talk, we will give a new algorithm, to define individual's delay in acknowledging an external stimulus, i.e., the delay of Event Related Potential P300. An accurate measurement could lead to various applications including diagnosis of the state of brain and prediction of its vulnerability to diseases like dementia or Alzheimer.

Fuzzy Set Theory in Information Technology

Sansanee Auephanwiriyakul
IEEE Senior Member, Head of Computer Engineering Department, Chiang Mai University
sansanee@eng.cmu.ac.th

Abstract

Fuzzy set theory introduced by LotfiZadeh in 1965 is an extension of the classical set. It is a way to understand or explain uncertainty occurring in everyday life. Since then there are many theories and applications developed based on Fuzzy set theory. One of the areas that Fuzzy set theory has an impact on is the area of information technology, especially, data mining. One of important parts in data mining is classification or sometimes called decision making. Fuzzy inference system has been used as a decision making tool in several applications including medical image processing. In particular, it is used as a classifier in microcalcification detection in Mammogram. Recently, there is several extension of Fuzzy set theory including interval Type-2 fuzzy set theory. One of the major developments of the interval Type-2 fuzzy set theory is interval Type-2 fuzzy logic, an extension of fuzzy inference system. It is applied into medical image processing as well. There is also another popular area in Fuzzy set Theory, i.e., a theory on fuzzy vector manipulation. Many research groups have been working on developing a theory of function manipulation. In particular, a theory of how the function will perform if the function inputs are fuzzy vectors not vectors of number. The major area in this case is developing a classifier with fuzzy vectors as inputs and output(s). In this talk, these classifiers and their real applications such as an application in communication assessment will be presented as well. Another area of fuzzy vector manipulation is an area of information fusion when input information is not numbers but fuzzy numbers. This talk will also show the proposed information fusion applied in a military application.

Session : Core Database Technologies and Data Mining Techniques

- Pattern Mining from Online Social Media
  Basabi Chakraborty, Takako Hashimoto
- Grid-based Image Morphing
  Porawat Visutsak
- Adeptness Associative Learning Method for Real-Time Cardiac Arrhythmia Detection
  Mohamed Ezzeldin A. Bashir, Dong Gyu Lee, Ibrahim Musa Ishaq, Makki Okasha, Ho Sun Shon , Keun Ho Ryu
- A Novel Mathematical Descriptive System for Human Body-Shape Representation
  Sukationg Phuphatana, Pirawat WATANAPONGSE

Pattern Mining from Online Social Media

Basabi Chakraborty[1], Takako Hashimoto [2]
[1]*Faculty of Software and Information Science, Iwate Prefectural University, Japan.*
basabi@iwate-pu.ac.jp
[2]*Faculty of Commerce and Economics, Chiba University of Commerce, Japan.*
takako@cuc.ac.jp

Abstract

With the rapid growth of Internet, Information and Communication Technologies various web based social networks are emerging at a fast pace. People interact with each other through on line social networks and their decision making in every sphere of life is influenced by those interactions. Efficient mining and analysis of online social data can provide assistance to people in different social needs like crisis management, reputation analysis, customer profiling and product survey. In this lecture, some of our research works in which data mining technologies are used for online social media data analysis, are presented.

1. Introduction

Online social websites are nowadays growing at an exponential rate. The vast amount of consumer data generated in online social media provides tremendous challenges to researchers and analysts, who are trying to gain insights into human interaction and collective behavior. Researches are going on in developing efficient techniques and algorithms for analysis of online social media data from blogs, microblogs, websites like *Facebook, Digg, Twitter* etc. As a result, new applications related to economy, marketing, education, business or medical science are developed.

Topic extraction and modelling, crisis management, reputation analysis, customer profiling, product survey, opinion or sentiment analysis are some of the emerging areas of popular applications of social network data analysis.

In this lecture, following three research works are presented.

1) Topic extraction from buzz marketing websites for analysis of the trend of market and reputation of specific products [1].
2) Analysis of victim's needs after a disaster from social media blog for developing better crisis management system [2].
3) Discovery of transition of people's concerns after a disaster from various social media to explore the dynamics of social needs' change [5][6]

In the next section, the first work on reputation analysis of market products from consumer purchasing support site is explained. In the following section, analysis of social needs and its transition over time from various online social media has been explained. In the 4th section, evolution of social dynamics after disaster event, analyzed from Facebook data has been represented followed by conclusion section.

2. Product Reputation Analysis from Websites

In 1) user messages in marketing web sites containing opinion of users about market products are analyzed using text mining techniques. Reputation analysis of the products are done by clustering topics in the users' messages. We used "kakaku.com" website for our experiment. It is the most popular "customer purchasing support site" in Japan. We selected messages for the five types of models of electronic air cleaners and collected 702 messages containing users' comments. We proposed the following frame work for data analysis: The main steps are:

1) Topic extraction by clustering: Here we extracted important keywords with high tf-idf value and used them as features for manually clustering the data into groups. We got 7 topic clusters as a preliminary result.
2) Important topic detection: We defined a parameter *contributing rate* to express the importance of a message. Contributing rate of topic A is defined as the ratio of topic A to all topics. We rank the topics according to their importance
3) Emerging needs detection: In this stage we tried to identify the topics associated with high

positive/negative emotion. Using a dictionary, we calculated the emotional degree of a topic by the positive /negative values of the words in the topic. Topics with high emotional degrees are considered to be the candidate of topics with emerging needs.

4) Visualization: In this stage, we used some visualization technique for validation of the cluster results.
5) Refinement: In the final stage, we defined a parameter *cluster quality index* to determine the correct value of number of topic clusters. *Cluster quality index* is defined as the ratio of intra-cluster distance to inter-cluster distance for a specific number of clusters. The smaller the *index*, the better the cluster result. Using this index we determined the most appropriate number of topic clusters.

3. Social Needs' Transition Analysis after Disaster

After the Great East Japan Earthquake occurred on March 11th, 2011, triple disasters (earthquake, tsunami and nuclear plant accident) crippled East Japan's (Tohoku) regular life. People heavily used social networks to provide and receive information and exchange opinions and sentiments. In 2) and 3) we analyzed several online social media data to find out victims' needs and how the need of the affected people changed with time. The results of our analysis can be efficiently used for development of efficient crisis management system or predicting the dynamics of human behavior and change of social needs after a disaster.

In a series of research works [2] [3] [4] [5] [6] we developed several technique for analysis of various social networks data like video streaming sites, blog posts, etc. after the Great East Japan Earthquake related disasters. Our basic framework of analysis are as following:

1) Data crawling
2) Language processing
3) Topic extraction
4) Time series topic detection
5) Visualization

Steps 1 and 2 are the part of preprocessing in which data from online social media is collected, words are extracted by morphological analysis (Japanese morphological analyzer Mecab is used), and the score of an individual word in a document is calculated by RIDF (residual IDF) measure. The words having high RIDF values are selected as keywords. In the next step we used different techniques for topic extraction and visualization of topic transition over time.

3.1 Graphical Method with Clustering [3]

In this method graph networks are constructed by the related words from keyword list. Related words are obtained using co-occurrence frequencies in a document. Graph networks are then clustered using a newly defined modularity value to decompose graphs in hierarchical modular structure corresponding to hierarchical structure of emerging topics. The similarity of the topics over time scale is calculated by considering the overlapping of key words. The correlation between two similar topics over adjacent time zones are calculated by Matthew's correlation coefficient. When the calculated value exceeds a threshold, it denotes the transition of topics over time. Video streaming sites and kakkaku.com web sites are used for data collection and the transition of topics are noted after the East Japan Earthquake.

3.2 Topic transition detection by Latent Semantic Analysis [6]

In another work, the keyword list from the preprocessing step is used for preparing a document-term matrix. The rows correspond to the documents and columns correspond to terms, the keywords with high RIDF score. Latent semantic analysis technique is used for analyzing relationship between document and term. Relationship contain a set of concepts related to document and term. By using LSA technique, the hidden concepts termed as topics here, can be derived from document-term matrix. To find tropic transition trend, tensor analysis of document-term-time tensor is needed. Tensor decomposition algorithms are necessary for that. To reduce the computational burden, as an approximation, we manually investigated the content of all hidden topics by its characteristic words. The topic types were used as tags to hidden topic types and manually evaluated the transitions of hidden topic types.

We used NPO Save Iwate's blogs and several BBS in the affected area to collect our data. From the blog, we analyzed data from 3rd June to 28th Dec December, 2011, 150 messages, and found 12 topics with a characteristic transition of each of them. From the results we could assess the victim's needs just after the disaster and the change of needs over time. This findings has great implication in developing better crisis management system in future. The change of needs of affected people also shows us the importance

of different social needs that should be offered to the people at different times.

3.3 Topic Transition Analysis with LDA

Recently LDA (Latent Dirichlet Allocation) technique has been used for topic transition analysis. LDA is a widely used multi topic document model based on Bayesian inference method. The original algorithm is computationally heavy so we used a simple version by R programming package. In LDA, the number of topics are to be set previously. We tried to set it from our previous experiments. We used NPO Save Iwate's blog data for the period of Jun 11 to July 12 and detected month wise topic/need's transition over this time. We then showed our results to NPO people and tried to refine the results taking account of their feedback.

4. Modeling Social Dynamics after Disaster [7]

In this work we tried to model social awareness and how the emotion of people flows after any disaster event by use of computational intelligence techniques. We extracted data from Facebook and Blog messages after terror strikes in Mumbai, India. The social dynamics generated by the behavior and flow of emotion, sentiment and opinion through social networks was modeled by ant colony behavior and swarm intelligence is used for determining the outflow of emotion and opinion. The interesting observations were found which depicted the behavior of transient crowd. We used the result to make conclusion about social awareness from the evolving social dynamics.

5. Conclusions

In this lecture, analysis of online social network data has been presented. Online social network data contains various information regarding the emerging needs of people, social dynamics, changing needs of affected people after any disaster, opinion and sentiment of people regarding market products to any sort of social perturbation. Online social networks are gradually becoming society's mirror and can provide great opportunity to influence society as a whole. The mining of evolving patterns of society from online social media is interesting and has a great effect in building better management system in case of disasters.

6. References

[1] Takako Hashimoto, Basabi Chakraborty and Yukari Shirota, "Social Media Analysis- Determining the Number of Topic Clusters from Buzz Marketing Sites", *International Journal of Computational Science and Engineering* (IJCSE), Vol 7., No. 1. pp. 65–72, 2012.

[2] Takako Hashimoto, Basabi Chakraborty and Yukari Shirota, "Topic Transition Detection about the East Japan Great Earthquake based on Emerging Modularity over Time", *Int. J. Computational Science and Engineering (in Press).*

[3] Takako Hashimoto Tetsuji Kuboyama, Basabi Chakraborty and Yukari Shirota,"Discovering Topic Transition about the East Japan Great Earthquake in Dynamic Social Media", *Proc. of IEEE International Conference on Global Humanitarian Technology (GHTC2012)*, 2012, pp. 259–264.

[4] Takako Hashimoto Tetsuji Kuboyama, Basabi Chakraborty and Yukari Shirota, "Discovering Emerging Topic about the East Japan Great Earthquake in Video Sharing Website, *Proc. of IEEE TENCON 2012*, Cebu, Philippines.

[5] Tatsuya Suzuki, David Ramamonjisoa, Basabi Chakraborty and Takako Hashimoto, "Extracting and Visualizing Peoples' Needs and Topics Trends from Users' comments on Video streaming or blog posts", presented in *SIG-FPAI in Morioka* on 28th Feb, 2013, appeared in FPAI B204, 2012, pp.13–18.

[6] Takako Hashimoto, Basabi Chakraborty, Tetsuji Kuboyama and Yukari Shirota,"Temporal Awareness of Needs after East Japan Great Earthquake using Latent Semantic Analysis", *Proc. of EJC2013 (23rd European-Japanese Conference on Information Modelling and Knowledge Bases)*, June, 2013, pp.214–226.

[7] Basabi Chakraborty and Soumya Banerjee, "Modeling the evolution of post disaster social awareness from social web sites", in *Proc. of the IEEE CYBCONF2013*, Laussane, Switzerland, June, 2013.

Grid-based Image Morphing

Porawat Visutsak
Department of Information Technology, Faculty of Industrial Technology and Management,
King Mongkut's University of Technology North Bangkok
porawatv@kmutnb.ac.th

Abstract

Image morphing is often used in film and television industry to produce synthetic visual effects by depicting the continuous evolution between two digital images. This paper presents a new method of 2D image morphing based on grid transformation. A grid is a series of intersecting horizontal and vertical lines that serve as guides to specify the coordinates of control points, or landmarks. Grid-based image morphing achieves a smooth transformation by incorporating transition of grid to maintain geometric alignment throughout the metamorphosis process. The proposed method has been successfully tested with a wide variety of images. The resulting sequence of images is better in visual quality and faster in execution time.

1. Introduction

Image morphing is a technique in film and television industry which has been developed for edutainment purpose [1, 2] for the last decade. It involves the smooth transformation of the objects and the colors of one digital image to the other [3]. The 2-D morphing problem concerns with the deformation of a source image to a target image [4]. The term "morphing" means warping two images with color interpolation. Image warping applies 2D geometric transformations on the images to retain geometric alignment between their features, while color interpolation blends their color [5].

This paper presents a new morphing technique: Grid-based Image Morphing. This method produces the smooth transformation of animated morphing images. By locating a grid as the control structure of image, together with a smart transition of grid, the new method can maintain geometric alignment during the morphing process. To perform a smart transition, the method computes the best-move of a grid in each frame corresponding to the number of desired frames for animating the transformation and the location of a grid of the source image and the target image. The proposed algorithm has been implemented and tested on different types of source and target images. Figure 1 shows the example of results. The experimental results demonstrate that the proposed algorithm produces results better in visual quality and faster in running time.

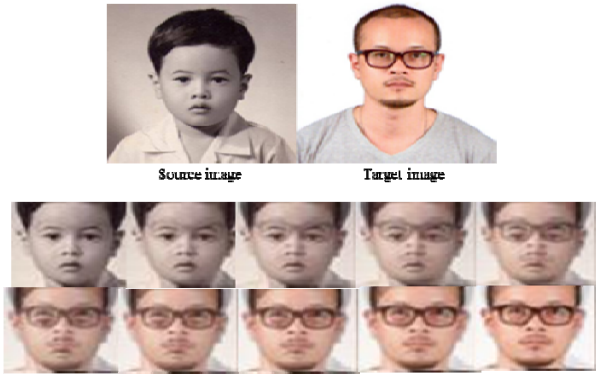


Figure 1. Example of animated transformation in the proposed method: The source image will be viewed as it is changing slowly frame by frame to the target image (Total execution time < 2 seconds per 10 frames)

The rest of the paper is organized as follows: Section 2 discusses about some existing techniques with their shortcomings. Section 3 describes the proposed method: Grid-based Image Morphing. Working procedure of the proposed technique is clearly described with some illustrations in this section. Section 4 shows how an image is being changed while it has been morphed; the evaluation of the proposed technique is also presented in this section. The conclusion and the future discuss are drawn in section 5.

2. Related work

A fair amount of several morphing algorithms have been published. Some of them are discussed in this

section including one-dimensional morphing, cross-dissolve morphing, mesh warping, and field morphing.

2.1 One-Dimensional Morphing

This simplest use of the technique to animate morphing among two images was introduced by [6, 7]. This method uses only one input and two images for the approximation network which can be thought of as degree of morph. Figure 2 is an example of manually detected feature points from source image to target image. Both images contain 26 related feature points.

Figure 2. 26 feature points, and manually detected feature points (white dots) in source and target images, respectively

The RBF (Radial Basis Function) network is structured with only one input I (degree of morph), two Gaussian bells in HL (hidden layer), and 52 OL (output layer) neurons (x and y coordinates of 26 features points), as shown in figure 3.

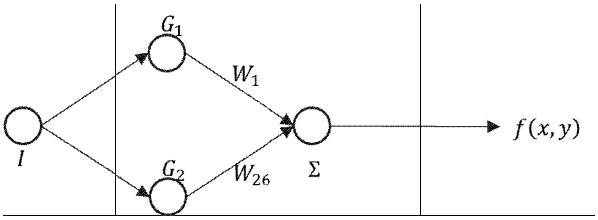


Figure 3. One-dimensional morphing RBF network with one input, two hidden layer neurons, and 52 outputs.

The degree of morph is set to be in the range of [0, 1]. At the learning stage, the RBFs network maps from input I = [0, 1] to corresponding training image feature points. At the second stage, the trained RBF network generates in-between images according to a given new input vector Inew = [0 0.333 0.667 1]. These in-between images (4 frames) are shown in figure 4. Frame no.1 and frame no.4 (source and target images) are originally given as training images, and frame no. 2 and frame no.4 are generated by RBF network. The second image, obtained by a degree of morph 33.3% which means that the image as 33.3% source image and 66.7 target image. More works on one-dimensional morphing were developed using SVMs (Support Vector Machines) instead of using RBF network [8, 9].

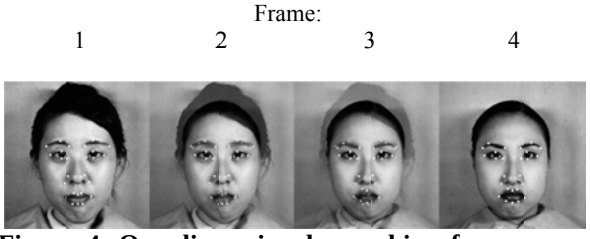


Figure 4. One-dimensional morphing from source to target images

2.2 Cross-Dissolve Morphing

Unlike one-dimensional morphing, cross-dissolve morphing does not determine the feature points detection from source image to target image. Cross dissolve morphing transforms one image into another image using linear interpolation. This technique is visually poor because the features of both images are not aligned, and that will result in double exposure in misaligned regions. In order to overcome this problem, the additional method called image warping is used to align the two images before cross dissolving. Image warping controls the color transition in the intermediate images produced. To morph one image to another, image warping determines the way pixels from one image are correlated with corresponding pixels from the other image. New positions and color transition rates for the pixels in each of the images in the sequence must be calculated [10]. The result of cross-dissolve morphing is shown in figure 5.

Figure 5. Cross-dissolve morphing

2.3 Mesh Warping

Mesh warping or Mesh morphing compute the deformation of the mesh from the source to the target domain based upon interpolation and/or extrapolation of the vertex coordinates. Mesh of the source image specifies the coordinates of control points, or landmarks. The second mesh of the target image specifies their corresponding positions in the target image [10]. The meshes are overlaid on source and target images. Feature points selection such as the eyes, nose, and lips lie below corresponding grid lines in

both meshes. The spatial transformation for mapping all points in source image onto target image is defined later on using the meshes. Figure 6 shows the deformation of mesh using warping method.

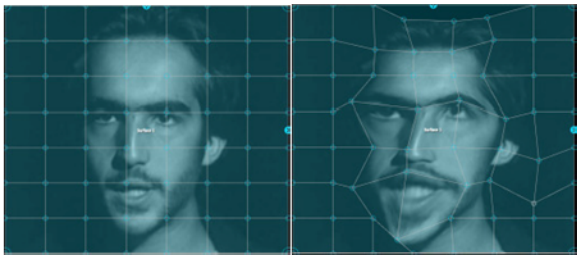


Figure 6. The deformation of mesh using warping method (Source images: http://1024d.wordpress.com/tag/tutorial/, last retrieved 14.06.2014)

2.4 Field Morphing

Field morphing [11] is more expressive and intuitive than mesh warping. This method uses the specification of feature points as lines, line features have been specified in both source and target images. For every intermediate position in morphing sequence, a line feature set is generated by interpolating the two sets between source and intermediate line feature sets. Every pair of line features represents a coordinate transformation for a point from source image to target image. A warp can be considered using the total warp function (a weighted sum of displacements due to all line pairs give net displacement of a point), then both source and target images have been warped to get two intermediate images. The color interpolation is then used to obtain the morph image. The field morphing algorithms is repeated for every position in the sequence to obtain the morph sequence.

3. Proposed Method: Grid-based image Morphing

Based on 10 feature points, which are 2 positions of eyes, 2 positions of mouth, and the other 6 conture points of the face (which are the 6 intersection points of eyes and ending points of mouth), the method defines the grid-partition as shown in figure 7.

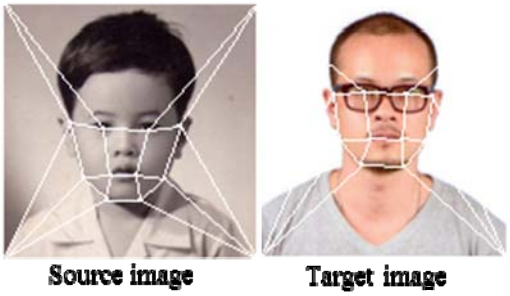
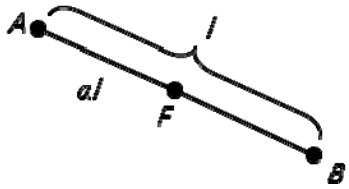


Figure 7. Grid partitioned of source and target images, respectively

Since the feature points of source and target images are at different positions, the images have to be warped in order to match their feature points. Figure 8 shows the morphed image (all 10 feature points must be matched; otherwise, the morphing process will generate the result image with too many eyes, noses, and mouths).

Figure 8. Warped image and matched 10 feature points

Next, in order to generate the sequence of animated images between source and target images, the weightings parameters for source and target images are assigned which are alpha and 1-alpha, respectively. For a feature point A in source image, and the corresponding feature point B in target image, by using linear interpolation, the position of the new feature point F would be generated as follow:



$$F = \alpha A + (1 - \alpha)B$$

The derived equation of F has been presented by [12], and the new feature point F is used to construct a point set which partitions the image in another way

different from source image and target image. Source and target images are warped such that their feature points are moved onto the same positions of all new feature points, and thus their feature points are matched. In the warping process, coordinate transformations are performed for each of the 16 regions respectively.

Since there exist many coordinate transformations for the mapping between two triangles or between two quadrangles (triangles and quadrangles within the grid). This study uses the affine and bilinear transformations for the triangles and quadrangles, respectively. In addition, bilinear interpolation is performed as pixel operation.

An affine transformation is a linear mapping from one triangle to another [12, 13]. For every pixel p within triangle ABC, assume the position of p is a linear combination of A, B, and C vectors. The transformation from triangle ABC to triangle DEF is given by the following equations,

$$\begin{gathered} p = \beta 1A + \beta 2B + \beta 3C \\ \beta 1 + \beta 2 + \beta 3 = 1, \beta i >= 0 \\ q = T(p) = \beta 1D + \beta 2E + \beta 3F \end{gathered}$$

where T() is transformation function

Here, there are two unknowns, β1 and β2, and two equations for each of the two dimensions. Consequently, β1 and β2 can be solved, and they are used to obtain q (the affine transformation is a one-to-one mapping between two triangles).

The bilinear transformation is a mapping from one quadrangle to another [12, 13]. For every pixel p within quadrangle ABCD, assume that the position of p is a linear combination of vectors A, B, C, and D. Bilinear transformation from quadrangle ABCD to quadrangle EFGH is given by the following equations,

$$\begin{aligned} p = {} & (1-u)(1-v)A + u(1-v)B + uvC + (1-u)vD, 0 <= u, v <= 1 \\ q = {} & (1-u)(1-v)E + u(1-v)F + uvG + (1-u)vH \end{aligned}$$

So, u and v can be solved (2 equations with 2 unknowns), and they are used to obtain q (the bilinear transformation is a one-to-one mapping for two quadrangles).

As a result of the coordinate transformations for each of source and target images, the feature points of these images are matched. As seen on figure 8, all 10 feature points in the source image will be at the same positions as in the target image. To complete morphing process, the proposed method has implemented cross-dissolve algorithm. Cross-dissolve method is described by the following equation,

$$C(x,y) = \alpha A(x,y) + (1-\alpha)B(x,y), 0 <= \alpha <= 1$$

where A,B are source and target images, and C is the morphing result.

4. Experimental Result

The study uses this method to build the sequence of morphing based on grid transformation and cross-dissolve algorithm. As the results are shown in figure 1, the morphing software constructed the sequence of 10-frame images morphing from source to target images with the execution time 1.35 seconds. This approach produces the smooth transformation of animated morphing images. As seen on the mid-way of morphing process, the transition of source image towards target image is smooth, and it is little clear but seems like another image is rising up from the background. Experiments reveal that the proposed algorithm gives good results and yields the good run-time in most PC environment.

5. Conclusion

This paper presents a very simple but important matter in image morphing. The study is useful to morph between two still images. The grayscale images are used in this study. The results of animated morphing sequence are smooth frame by frame. The experiments reveal that the algorithms based on grid transformation give good morphing quality run-time and the smooth transition during the morphing process. In the future, the further study of grid transformation for image registration might be investigated and the implementation in the other image applications must be surveyed.

Acknowledgment

"This research was funded by the King Mongkut's University of Technology North Bangkok. Contract no. KMUTNB-GEN-57-29".

7. References

[1] S. M. Seitz, "Bringing photographs to life with view morphing", *Proc. INA Imagina 9*7, 1997, pp. 153-158

[2] Wolberg, G, "Image morphing: a survey", *The Visual Computer*, Volume 14, Issue 8-9, 1998 , pp 360-372

[3] M.Shahid Farid, Arif Mahmood, "Image Morphing in Frequency Domain", *J Math Imaging Vis*, 42, pp. 50-63

[4] J. Chalidabhongse and C.-C. Jay Kuo, "A Multiresolution Approach for Image Morphing". *Signals, System and Computers,* Vol.1, 1993, pp. 16-20

[5] Wolberg, G., "Recent Advances in Image Morphing", *Proc.Computer Graphics International 1996,* 1996, pp.64-71.

[6] Porawat VISUTSAK and Korakot PRACHUMRAK, "The Skeleton Pruning-Smoothing Algorithm for Realistic Character Animation", *JMMT: Journal of Man, Machine and Technology*, Vol. 2, No. 1, 2013 , pp. 21 ~ 34

[7] Porawat VISUTSAK and Korakot PRACHUMRAK, "Geodesic-based Skeleton Smoothing", *International Journal of Mathematical Models and Methods in Applied Sciences,* Issue 4 Vol.5, 2011, pp. 713-721

[8] Porawat Visutsak, "Emotion Classification through Lower Facial Expressions using Adaptive Support Vector Machines", *JMMT: Journal of Man, Machine and Technology*, Vol. 2, No. 1, 2013, pp. 12 ~ 20

[9] Porawat VISUTSAK, "Emotion Classification using Adaptive SVMs", *International Journal of Computer and Communication Engineering*, Vol.1, No.3, 2012, pp. 279-282

[10] Rahman, Md Tajmilur, et al., "A novel approach of image morphing based on pixel transformation", *Computer and information technology, 2007. iccit 2007. 10th international conference on. IEEE, 2007,* 2007

[11] Nishita, Tomoyuki, Toshihisa Fujii, , Eihachiro Nakamae., "Metamorphosis using Bezier clipping.", *Proceedings of the First Pacific Conference on Computer Graphics and Applications.,* 1993

[12] Jonas Gomes et al., "Warping and morphing of graphical objects" , *Morgan Kaufmann Publishers* , 1999

[13] Martin Bichsel, "Automatic Interpolation and Recognition of Face Images by Morphing", *Proceedings of the 2nd international conference on automatic face and gesture recognition*, 1996, pp128-135

Adeptness Associative Learning Method for Real-Time Cardiac Arrhythmia Detection

Mohamed Ezzeldin A. Bashir[1], Dong Gyu Lee[2], Ibrahim Musa Ishag[2], Makki Okasha[2], Ho Sun Shon[2], Keun Ho Ryu[2]
[1]*University of Medical Science and Technology, Sudan*
{mohamed.izz}@umst-edu.net
[2]*Database/Bioinformatics Laboratory, Department of Computer Science,*
Cheungbuk National University, South Korea
{dglee, ibrahim, makki,shon0621, khryu}@dblab.chungbuk.ac.kr

Abstract

It is vital for the automated system to accurately detect and classify ECG signals very fast to provide a useful means for tracing the heart's health in the real time. Making a random training set can lead or cause negative results. Simply, training set must designed carefully to consider all possible classes of the overall arrhythmia, so as to train the algorithm with the right group. Not only, considering all possible arrhythmia, but also with the same ratio, which means giving the algorithm a richness group to be trained in the right way with effective training set. Therefore, a means is needed for determining which record in the main file to be selected to replace the removed one from the active training data set. Our proposed methodology automatically trains the classifier model, using efficient set. Accordingly the experimental works show an improvement in the performance of some classification models to detect cardiac arrhythmia.

Keywords: Electrocardiogram (ECG); Arrhythmia; Classification; and Training dataset

1. Introduction

ECG signals are a very important medical instrument that can be utilized by clinicians to extract very useful information about the functional status of the heart. So to detect heart arrhythmia, which is the anomalous heartbeat map with a different shape in the ECG signal notice by deflection on the P, QRS, and T waves, some parameters are acquired and an enormous finding is produced [1].

The ECG gives two major kinds of information. First, by measuring time intervals on the ECG, the duration of the electrical wave crossing the heart can be determined, and consequently we can determine whether the electrical activity is normal or slow, fast or irregular. Second, by measuring the amount of electrical activity passing through the heart muscle, a pediatric cardiologist may be able to find out if parts of the heart are too large, or overworked. Electrocardiography has evolved over time and is becoming more accurate as it is being automated by making use of several software techniques available.

There has been a great deal of interest in systems that provide real time ECG classification through an intermediary local computer between the sensor and control center [2]. It is vital for the automated system to accurately detect and classify ECG signals very fast to provide a useful means for tracing the heart's health in the right time. The effectiveness of such systems is affected by several factors, including the ECG signals, estimated ECG's features and descriptors, the dataset used for learning purpose, and the classification model applied [3].

This paper, concerned with the challenges for training the classifiers model with updated data to facilitate the process of developing real time cardiac health monitoring systems. It presents a method that propos solution to solve this problem. In the rest of this paper, we will provide a brief description of related work, associative learning method, and then we present the experimental works and finally the conclusion.

2. Arrhythmia Detection Training Methods

Confirming local and global dataset are the main two approaches of training dataset used to learn the classifier model. Global is built from a large database that most automatic ECG analysis research works refer

to this technique such as [3], [4]. Simply, there are training and testing datasets with different percentages through which the classifier is trained using the training dataset, and later predict the unseen group of data through the testing dataset. (Rodriguez, J. et) attempted to derive approach that can build the most accurate model for classifying cardiac arrhythmia based on feature extraction [5]. He divided the dataset into random groups one for training (66%) and another for validation (33%). He used the "weka" and "answertree" tool in his experiment. Sixteen methods were used in the experiments. One main challenge faced by this technique is the morphologies of the ECG waveforms that widely vary from patient to patient. Accordingly, the classifier learned by specific data related to an identical patient will perform very well when tested with unseen data of that patient often fail when presented with ECG waveforms for other patients. To overcome this problem, the literature shows that there is a trend to learn the classifier via training dataset as much as possible. This was the commercial trend introduced by the ECG device vendors. However, such an approach criticized in different aspects. First, when using a huge amount of ECG records to build a classifier, development, maintaining, and updating will become very complicated. Second, it is difficult to learn the classifier by abnormality ECG during the monitoring process. Therefore, there is the possibility to be unable to detect specific arrhythmia when applying that model to patient records. Moreover, it is impossible to introduce all ECG waveforms from all expected patients [6].

In previous works, we suggested a nested ensemble technique to solve the problem of training dataset by manipulating the training dataset for learning the classifier through up-to-date data, and manipulating the ECG features to select the proper adequate set (morphological features) to enrich accuracy [7]. Although the results are favorable, synchronizing the two components is expensive, which negatively affects the detection of the arrhythmia in real time. Moreover, it is quite static to some extent. Then Trigger Learning Method [8] and Active learning method [9] have been suggested to detect cardiac arrhythmia on line in very sufficient manner; simply introduced to learn the classifier model by up-to-date training data.

The local learning set is a customized to a specific patient; in other words, it is a technique focused on developing a private learning dataset corresponding to each patient [10]. Its intention is to familiarize the classification model with the unique characteristics of each patient. Although this technique looks to alleviate the problem of the learning process, it suffers from a clear problem related to the difficulties to distribute an ECG database because it is time consuming and labor intensive. Moreover, few patients are accepted to be involved in the development of the ECG processing method. Thus, there are limitations to the advantages provided by such technique among the expected audience, even if it is permissible.

3. Associative Learning Method

The associative learning method comes to solve the problem of feeding the arrhythmia detection algorithm with updated training sets. The associative technique has four steps as shown in Figure 1. First, an initial learning stage is introduced to learn the classifier by a random set of data without any further consideration. The classifier performance is evaluated (check) and updated (improve) for consistency, and applied the re-movement stage to avoid a combat situation. In the initial learning step, the learning process starts by utilizing a random group of records (categories), which represent (50%) of the overall dataset. In the check stage, the overall trusted mark, which is calculated using the local trusted mark that can be measured using a label assigned to the specific category with a specific vector of features.

$$L^M(x) = \sum_{f \in features} \beta_f(F,i).C^S(x) \quad (1)$$

where f is the feature number, F is the contribution of the feature, and CS(x) represents the category score when labeled as arrhythmia (i), which calculated as follows:

$$C^S(x) = \sum_{f \in features} \beta_f(F,i) \quad (2)$$

The function βf (F, i) checks the set of features (F) in specific category labeled as arrhythmia (i). It returns "+1" if the label (i) is assigned to category (x), otherwise it returns "−1."

The local trust index LM(x) is considered in determining the overall trust index TrustM(X), which is defined using a sigmoid function Sigmoid(X) ($0.5 < TrustM(X) < 1$).

The checking step ends by two judgments; either the current training set is reliable or not, depending on different classes of arrhythmia assigned to categories. Accordingly, the unreliable set needs to be modified by a new group of data. This process has two parts, first, specifying the useless category or categories; and second, replacing it or them with newly selected one(s). In the first step, category (x) in the active training set (X) is removed if the category score CS(x) is less than

a threshold δremove. Second, for the record selection the method introduce in-between set of data file to avoid delay, which is called cache the cache is used to substitute the partial or complete modification of the current active training data set. The objective is to minimize hitting the main database as much as possible, so as to save time and increasing the accuracy by double filtering. Starting from the second modification stage, the substitutions take place from the cache not from the main database. The improvement step is active when there is a limited number of bad labeling using the current group, while it is useless when there are multiple defects among the categories the thing that requires an inherited improvement process consequence. It is very expensive in terms of time, which negatively affects the performance of the classifier model to label different types of arrhythmia. Therefore, the re-movement step is introduced to deal with this problem.

The selection of the substituted category depends on record reference by permitting each record belongs to specific arrhythmia class to be selected when replacement is required. In this case, the cache control point interprets a recode reference simply as a tag and a feature set with arrhythmia class. The tag field uniquely identifies a recorded to determine whether it is in the cache or in the main file, the cache control point must simultaneously examine every line's tag for a match.

The removed category or categories will be sent to the cache not to the main database. Thus the removed categories will be sent to the cache with their record references. The cache size is fixed so as not to exceed 20% of the total dataset size.

Category will be removed from the cache if selected twice and removed from the active learning set. In this case, it will be replaced with a new category from the main database to avoid selecting the same one.

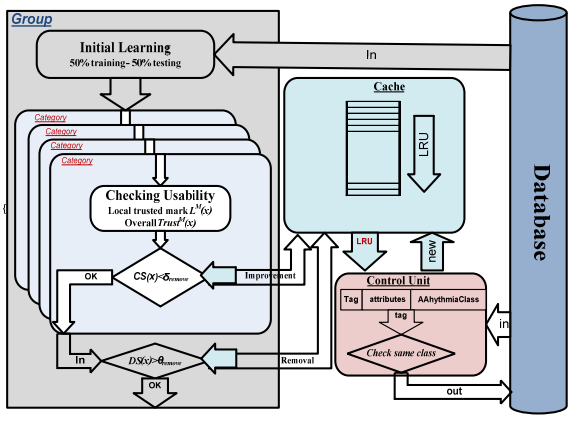


Figure 1. Associative learning process flow

4. Experimental Results

4.1. Environment

We used a database generated at the University of California, Irvine [11]. It was obtained from Waikato Environment for Knowledge Analysis (WEKA), containing 279 attributes and 452 instances [12]. The classes from 01 to 15 were distributed to describe normal rhythm, Ischemic changes (Coronary Artery Disease), Old Anterior Myocardial Infarction, Old Inferior Myocardial Infarction, Sinus tachycardia, Sinus bradycardia, Ventricular Premature Contraction (PVC), Supraventricular Premature Contraction, Left bundle branch block, Right bundle branch block, degree Atrio ventricular block, degree AV block, degree AV block, Left ventricle hypertrophy, Atrial Fibrillation or Flutter, and Others types of arrhythmia Respectively. Some instances related to specific arrhythmia classes are duplicated generating overall 573 instances. The experiments were conducted in WEKA 3.6.1 environment. Our experiment was carried out by a PC with an Intel Core processor (T M) 2 DUO, speed to 2.40 GHz. And RAM 2.00 GB.

4.2. Necessity of Including All ECG Parameters

First, we proved the necessity of including the P and T waves in conjunction with the QRS complex to evaluate arrhythmia the right way. We measure the performance of five different algorithms including OneR, J48, Naïve Bayes, Dagging, and Bagging according to the parameter(s) used to classify the arrhythmia. Table I summarizes the results obtained by each algorithm.

Table 1. The accuracy according to specific ECG parameter

Features	OneR	J48	Naïve Bayes	Dagging	Bagging
QRS only	60.4	91.2	76.5	63.5	81.0
QRS + P	60.4	91.4	77	62.4	81.6
QRS + T	61.3	91.2	76.7	63.0	82.3
QRS + P +T	61.1	92.3	77.7	64.2	83.0

4.3. Arrhythmia Detection

Figure 2 compares the accuracies achieved by the OneR, J48, naïve Bayes, dagging, and bagging methods when using the assciative learning. We also show their original performance without the proposed method for comparison.

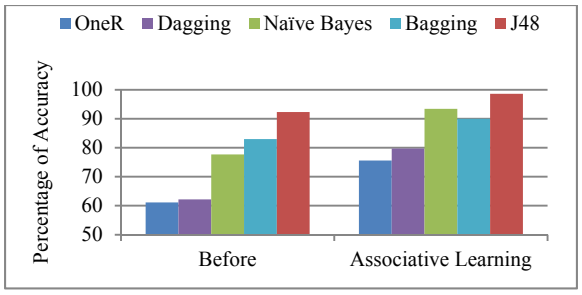


Figure 2. Accuracy achieved by different methods when using associative technique

Figure 3 illustrates the improvements due to the proposed associative learning in all algorithms tested here. We specifically compare the best-case accuracies when including all features related to the P, QRS, and T waves with that obtained after using the associative learning .

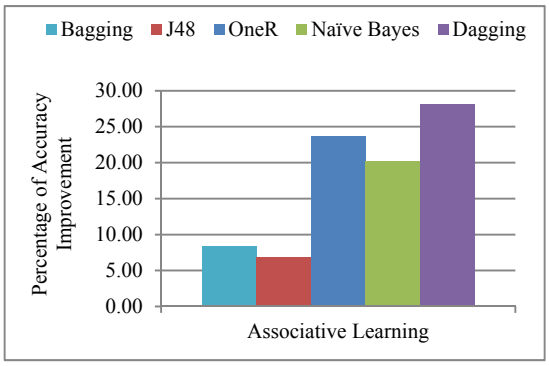


Figure 3. Accuracy improvement achieved by associative technique

Figure 3 clearly show that the associative learning improve the detection accuracy for the different types of arrhythmia. The improvement is noticeable for all algorithms with different weights due to their mechanisms. Specifically, improvements of 23.7, 6.8, 20.2, 28.1, and 8.4 percentages were achieved in performance for OneR, J48, naïve Bayes, dagging, and bagging, respectively, when applying the associative technique. In general, these are significant improvements.

It is also interesting to compare the accuracy of associative technique using the J48 algorithm with that of other methods presented in the literature. Methods from two representative studies were chosen for this comparison, which including trigger learning method [8] and active learning method [9]. Table II summarizes the comparative results of these methods, in which the last row lists the results of associative method. Among the two methods, the proposed method outperforms the other methods with an impressive accuracy of 98.6% in discriminating 15 ECG beat types.

Table 2. Accuracy comparison with other methods

Method	Accuracy %
Trigger learning method	96.1
Active learning method	97.6
Associative learning method	98.1

5. Conclusion

Cardiac health monitoring is a challenging problem in the field of data mining and knowledge extraction, and has received considerable attention over the past few years because of its importance in saving lives and reducing health risks. Today, cardiac health monitoring has reached a level of maturity when operating directly on or off-line. However, current methods are far from adequate for automated, remote cardiac health monitoring by detecting arrhythmia in real time. This is partly because of inter- and intra-patient variabilities. Thus, developing one classifier model to satisfy all patients in different situations using static training datasets is not practical. Furthermore, analyzing the QRS, P-wave, and other elements of ECG, and measuring the time interval between these elements, is necessary for real-time cardiac monitoring. This is technically infeasible with current systems because of computational limitations.

In this paper, we presented a associative technique as a proposed solution to solve these problems. The performance of our method was evaluated using various approaches, which demonstrate their effectiveness. In future, we plan to perform more experiments to account for interrelated ECG features.

Acknowledgement

This work was supported by the grant of the Korean Ministry of Education, Science and Technology (The Regional Core Research Program / Chungbuk BIT Research-Oriented University Consortium), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2010-0001732).

6. References

[1] Dale Dubin, MD, “rapid interpretation of EKG”, *6th edition, cover publishing co.*, 2000

[2] U. Rajendra, M. Sankaranarayanan, J. Nayak, C. Xiang, and T. Tamura, "Automatic Identification of cardiac health using modeling techniques: a comparative study", *Inf. Science. 178*, 2008, pp. 457–4582

[3] M. Ezzeldin A. Bashir , M. Akasha, D. G. Lee, Min Yi, K. H. Ryu , E. J. Bae, M. Cho, and C. Yoo, "Highlighting the Current Issues with Pride Suggestions for Improving the Performance of Real Time Cardiac Health Monitoring," *DEXA Bilbao, Spain, LNCS 2010.J*, 2010.

[4] G. Bortolan, I. Jekova and I. Christov, " Comparison of four methods for premature ventricular contractions and normal beats clustering", *Comp. Card.* , 2005, pp. 921–924

[5] G. Clifford, F. Azuaje, P. McSharrg, "Advanced methods and tools for ECG data analysis", *Artech house*, 2006.

[6] H. Palreddy and W. Tompkins, "A patient-adaptable ECG beat classifier using a mixture of experts approach", *B. med. Eng. 44*, 199, pp. 891–900

[7] M. Ezzeldin A. Bashir , M. Akasha, D. G. Lee, Min Yi, K. H. Ryu , E. J. Bae, M. Cho, and C. Yoo, "Nested Ensemble Technique for Excellence Real Time Cardiac Health Monitoring", *BioComp lasvegas USA 2010,* 2010

[8] M. Ezzeldin A. Bashir, D. G. Lee, M. Li, J. W. Bae, H. S. Shon, M. C. Cho, and K. H. Ryu, "Trigger Learning and ECG Parameter Tuning for Real-Time Cardiac Clinical Information System", *IEEE Transactions on Information Technology In Biomedicine 2012*, Vol.16 No.4, 2012, pp.561-571

[9] M. Ezzeldin A. Bashir, H. S. Shon, D. G. Lee, H. Kim, and K. H. Ryu, "Real-Time Automated Cardiac Health Monitoring by Combination of Active Learning and Adaptive Feature Selection", *KSII Transactions on Internet and Information Systems*., Vol. 7, No.1, 2013, pp.99-118

[10] Y. H. Hu, S. Palreddy, and W. J. Tompkins, Eds., "Patient adaptable ECG beat classification using mixture of experts", *in Neural Network for Signal Processing V. Piscataway, NJ: IEEE Press*, 1995, pp. 495–463

[11] UCI Machine Learning Repository, "http://www.ics.uci.edu/~mlearn/MLRepository.html".

[12] WEKA web site, "http://www.cs.waikato.ac.nz/~ml/weka/index.html"

A Novel Mathematical Descriptive System for Human Body Shape Representation

Sukationg Phuphatana, Pirawat WATANAPONGSE
Department of Computer Engineering, Kasetsart University Bangkok, Thailand
jojoth@gmail.com, Pirawat.W@ku.ac.th

Abstract

The Human body shapes are typically represented by the ordered treble Bust Waist Hip (BWH) measurement. Further classification of those shapes into body types classically employs the culturally biased "Apple, Pear, (Inverted) Cone, Hourglass, and Cylinder" descriptions. Even with the advent of the 3 D body scanner, those systems persist because of their simplicity and there are practically no other competing descriptive systems. This article proposes an equally simple, mathematically oriented, body shape descriptive system based on Geometric We use Polynomial Curve Fitting to represent cross section (top view) and side profile (front view) obtained from individual's 3 D body scan. The proposed system has the advantages of retaining more individual information without sacrificing its simplicity, and is also backward compatible with the classical systems.

Keywords: Body Shape; Body Shape Representation; Polynomial Curve Fitting; Geometric Curve Fitting; 3-D Body Scan

1. Introduction

In healthcare, Body type is used for workout and disease prevention. For workout, Pear shapes have lager bottom than top that means Pear shape has fat in hip thigh and butt more than other shape [23], so this shape should be exercise by focus on lower body than upper body to balance out body. For disease prevention, Apple shapes have more fat around abdomen that cause more risk for some conditions such as heart disease and diabetes than other shape. In fashion, Designer design cloth to match each shape .The right cloth to the right shape can improve your look instantly.

Classification of body type can do by Expert or general formula [24]. Expert often use one-view picture, generally front-view, of body for classification which cause multi-expert disagreement of body type in some cases. For non-experts, they use general formula [24] for classification but the formula doesn't have reference or concrete experiment result to support it. shape description has many techniques in the literature such as Manifold Representation[1], Topology Matching[2] and Volumetric[3].Oren Freifeld[1] define human shape as point on manifold, every point represents a deformation form template. That can measure via a geodesic distance (tangent space, vector space). M. Hilaga[2] present a technique, call Topology Matching, in which similar to polyhedral models by comparing Multiresolutional Reeb Graph.This technique describe shape in tree structure. As a result, that can't describe difference shape of body model in same posture. To descript difference shape, This technique often use for posture checking of human model. And Jigi Zang[3] proposes a shape signature, called Volumetric Extended Gaussian Image (VEGI). It captures the volumetric distribution of a 3D mesh model along the latitude-longitude direction without conventional pose normalization. This method isn't effected by translation and scaling. This technique is able to differentiate between non-convex and convex objects. This technique uniformly decomposes a 3D model into N concentric spheres. Which N for each model may not same N. JigiZang isn't describe method about find proper N for each model. However, all method above needs mesh data is input but raw data from 3d body scanner is point cloud and It need some preprocessing to build mesh. This article use simple mathematical equation to descript body shape by utilized side profile(side edge torso from bust to hip) and cross section of bust, waist and hip follow which base on BWS Measurement[10,11]. This has several advantages. First, this method is independence form translation, scaling and rotation. Second, this method can handle missing point of the edge of body model (number of missing point is much not more than thirty percent) Finally, This technique is not require mesh data; it can be applied to point cloud model directly.

2. Background

2.1. 3D Body Scanning

A 3d body scanning is a technology for the digitization of the human body. It enable to rapidly collect three-dimensional (3D) data (Figure 2) for a more efficient use of the resulting data such as mass customization, movie industry and manufacturers.

Figure 1. 3D body scanner

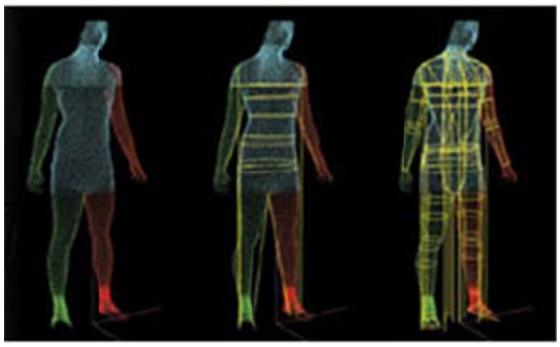


Figure 2. 3D data of 3D body scanner

2.2 Curve Fitting

Curve fitting is process fits equation of approximating curves to the raw series of data. Nerveless, for a given set of data, the fitting curves of a given type are generally not unique. Thus, a curve is a minimal deviation from all data. Different types of curve fitting [8]

2.2.1 Fitting Lines and Polynomial Curves to Data Points Polynomial equation is use to curve fitting.

2.2.1.1 First Degree Polynomial Function This is a line ith slop a. line will connect any two points, so a first degree polynomial equation is an exact fit through any two points.

$$f(x) = ax + b$$

2.2.1.2 Second Degree Polynomial Function This will exactly fit a simple curve to three points (quadratic polynomial).

$$f(x) = ax^2 + bx + c$$

2.2.1.3 Third Degree Polynomial Function This will exactly fit a simple curve to four points (cubic polynomial).

$$f(x) = ax^3 + bx^2 + cx + d$$

If there are more than $n + 1$ constraints (n being the degree of the polynomial), the polynomial curve can still be run through those constraints.

In general, the graph of a polynomial function of degree n has at most n x-intercepts, and at most n-1 turning points.

2.2.2 Fitting Other Curves to Data Points These types of curves, such as conic sections (circular, elliptical, parabolic, and hyperbolic arcs)

2.2.3 Algebraic Fit versus Geometric Fit for Curves For algebraic analysis of data, "fitting" usually means trying to find the curve that minimizes the vertical (y-axis) displacement of a point from the curve (e.g., ordinary least squares). However for graphical and image applications geometric fitting seeks to provide the best visual fit.

2.2.4 Fitting a Circle by Geometric Fit This method trying to find the best visual fit of circle to a set of 2D data points. The method elegantly transforms the ordinarily non-linear problem into a linear problem that can be solved without using iterative numerical methods

2.2.5 Fitting an Ellipse by Geometric Fit This method extended to general ellipses by adding a non-linear step, resulting in a method that is fast, yet finds visually pleasing ellipses of arbitrary orientation and displacement.

3. Data Acquisition and Processing

We use two view of 3d model, that is top view as cross section and front view as side profile .Our method base on two simple mathematic to descript shape representation(Figure3.). First method, we use Quaratic function with side profile to retrieve the polynomial function of side profile, then use Leading coefficient as a feature to identify shape representation .Second method we use Ellipse fit with cross section of bust ,waist and hip to retrieve perimeter of each. Then use three ratios as a feature to identify shape representation.

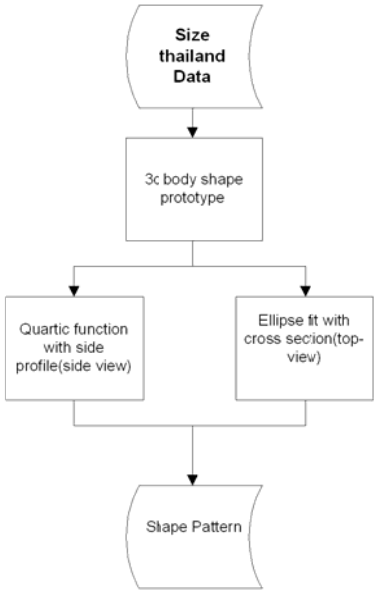


Figure 3. Overview of the process

3.1 Size Thailand Data

This is a sizing survey of the Thai population using 3d body scanning technology. That age between 16 to 59 years and data collected from 13,442 subjects from all parts of country.

3.2 3D Body Shape Prototype

We random 209 subjects from Size Thailand data and it are to three experts for classifying body type. We use only the data that got at least two expert agreement. We need at least fifty subjects for prototype. As a result we got three body types that have more than fifty subjects. That is Pear, Conical and Tube

3.3. Quartic Function with Side Profile

The first method we propose polynomial. For descriptor side profile (Figure4). We require 2 coordinate axis (e.g. x, z) side edge form bust to hip of human body (Figure 4) to calculate coeffient of non-liner polynomial, that is a 4th degree polynomial (Quartic function).We use 4th degree polynomial because this function has three turning points [21].First turning point represent curve of bust. Second turning point represent curve of waist. Third turning point represent curve of hip.

$$y = ax4 + bx3 + cx2 + dx + e \quad (1)$$

The first-order coefficient (Leading coefficient) is "a". We use sign of first-order coefficient. For describe behavior (Trend of polynomial graph), that is positive (Figure 5.) or negative.

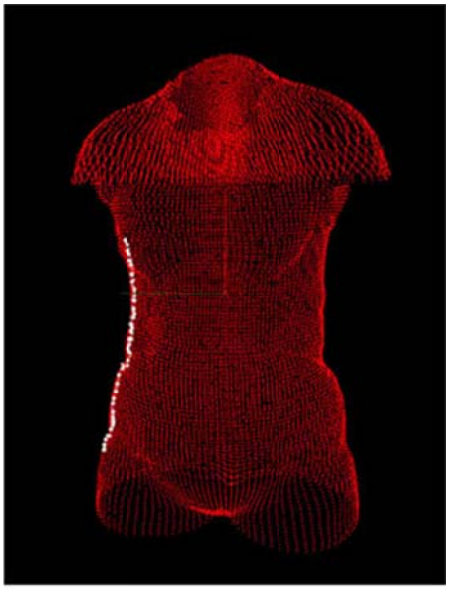


Figure 4. The white line that is side edge form bust to hip

TC^2 3d body scanner software provide hip coordinate but not bust coordinate, So bust coordinate we calculate from bust height. We use only torso of model.

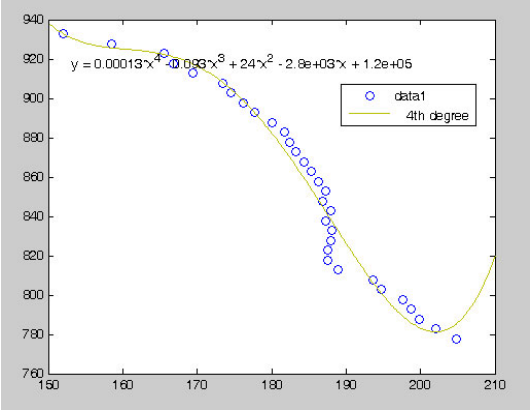


Figure 5. Show positive 4th degree Polynomial

3.4 Ellipse Fit with Cross Section

The second method of body shape descriptor. We use ellipse fit [22] to calculate perimeter of bust waist and hip because it produce minimal deviation of perimeter from cross section data.

$$ax^2 + 2bxy + cy^2 + 2dx + 2fy + g = 0 \qquad (2)$$

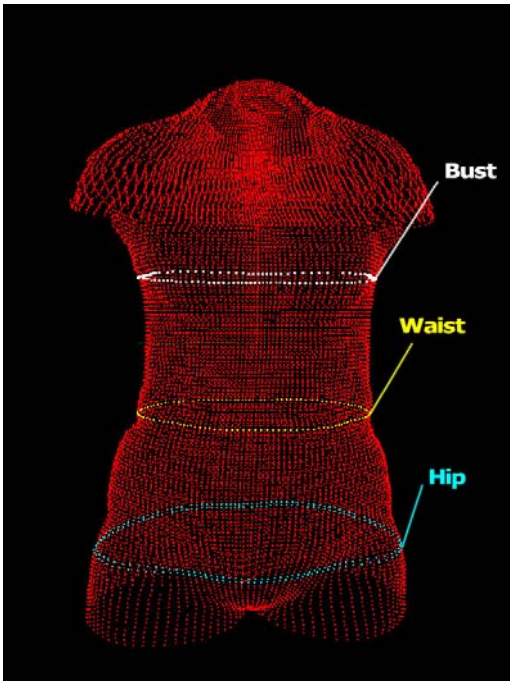


Figure 6. Show position of bust, waist and hip

We can calculate waist position by waist height from TC23d body scanner software

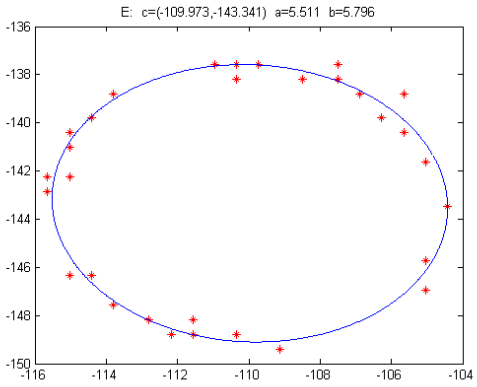


Figure 7. show ellipse with closed data with Ellipse Fit [9]

We calculate semi major/minor axis length (semi axis length)from formula(3) and use semi axis length calculate perimeter by formula (4)

$$a' = \sqrt{\frac{2(af^2+cd^2+gb^2-2bdf-acg)}{(b^2-ac)\left|\sqrt{(a-c)^2+4b^2}-(a+c)\right|}}$$

$$b' = \sqrt{\frac{2(af^2+cd^2+gb^2-2bdf-acg)}{(b^2-ac)\left|\sqrt{(a-c)^2+4b^2}-(a+c)\right|}} \qquad (3)$$

$$2\pi\sqrt{\frac{a^2+b^2}{2}} \qquad (4)$$

3.5 Shape Pattern

We use leading coefficient from 3.3 and three ratios from section3.4 to identify shape representation

4. Experimental Results

4.1Quartic Function with Side Profile

In Pear, 35models are negative and 15 models are positive. That is 70% of data are negative. Positive is consisting of ten shape expert and five shape global. In Tube, 41 models are positive and 9 models are negative. That is 81% of data are positive. In Conical, 48 models are positive and 2 models are negative.

4.2 Ellipse Fit with Cross Section

Pear has largest perimeter of hip. Conical has largest perimeter of bust. Those are the same with global description [10,11].Tube has average of three ratio as Bust-waist ratio is 1.136,Waist-hip ratio is

0.849and Bust-hip is 0.974. In Pear, Waist-hip ratio is1.081, Waist-hip ratio is 0.801and Bust-hip is 0.865. In Conical, Waist-hip ratio is1.133, Waist-hip ratio is 0.942and Bust-hip is 1.065

5. References

[1] Oren Freifeld, and Micheal J.Black, "Lie bodies: a manifold representation of 3D human shape," *Computer Vision –ECCV2012*, vol.7572, 2012,pp. 1-14.

[2] M.Hilaga,Y. Shinagawa,T. Kohmura,T.L. Kunii "Topology Matching for fully Automatic Similarity Estimation of 3D Shapes", *ACMSIGGRAPH* , 2001, pp.203-212

[3] Jigi Zhang,Hau-San Wong and Zhiwen Yu, "3D model metrieval based on volumetric extended gaussian image and hierarchical self-organizing map".

[4] Afzal Godil and Sandy Ressler, "Retrieval and Clustering from a 3D Human Database based on Body and Head Shape", *Digital Human Modeling for Design and Engineering Conference, FRANCE, Session: Advanced Size/Shape Analysis*, Vol 01, 2006, 2355

[5] http://www.sciencedaily.com/releases/2013/01/130110161350.htm

[6] Dr. Marie Savard, Apples & Pears: The Body Shape Solution to Weight Loss and Wellness

[7] http://www.sizethailand.org

[8] http://en.wikipedia.org/wiki/Curve_fitting

[9] http://www.mathworks.com/matlabcentral/fileexchange/22684-ellipse-fit--direct-method

[10] August, B. (1981). Looking thin. New York: Rawson Wade

[11] McCormack, Helen , "The shape of things to wear: scientists identify how women's figures have changed in 50 years". *The Independent. UK. How female body shapes have changed over time*, 2005

[12] Lihua Zhang ,Wenli Xu,Cheng Chang , "Pattern Recognition Letters", Vol 24, 2003, pp.9-19.

[13] Li,B.,Holstein,H., "Using k-d trees for robust 3D point pattern matching", *3-D Digital Imaging and Modeling*,2003, pp.95-102.

[14] http://en.wikipedia.org/wiki/Bust/waist/hip_measurements

[15] Marcel K.,Marcin N.,Reinhard K., "3D shape matching with 3D shape contexts ", *In The 7th Central European Seminar on ComputerGraphics*, 2003

[16] Levi C. Monteverde,Conrado R. Ruiz,Zhiyong Huang, "A shape distribution for comparing 3d models", *MM'07 Proceedings in13th International conference onMultimediaModeling-Volume Part I*, 2007, pp.54-63

[17] Christopher M. Cyr, Ahmed F. Kamal,Thomas B. Sebastian,Ben jamin B. Kimia, "2D-3D Registration Based on Shape Matching ", *MMBIA '00 Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 2000, pp.198

[18] Li B., Holstein H., "Using k-d tree for robust 3D point pattern matching ", *3-D Digital Imaging and Modeling*, 2003, pp. 995-102

[19] Hye won Seo and Nadia M., "An example-based approach human body manipulation", *Graphical Models*,Vol 68, 2004, pp.1-23

[20] Robert O., Thomas F., Bernard C. and David D, "Matching 3D Models with Shape Distributions", *SMI'01 Proceedings of the International Conference on Shape Modeling & Appplications*, 2001, pp.154

[21] http://mathworld.wolfram.com/QuarticEquation.html

[22] http://mathworld.wolfram.com/Ellipse.html

[23] http://www.sciencedaily.com/releases/2013/01/130110161350.htm

[24] McCormack, Helen, "The shape of things to wear: scientists identify how women's figures have changed in 50 years". *The Independent. UK. How female body shapes have changed over time.*

Session : Communication & Networking

- Three-dimensional image processing using Integral Imaging
  Ganbat Baasantseren
- Mining Frequent Itemsets in Transactional Database by Reduction of Trees Traversal
  Supatra Sahaphong, Gumpon Sritanratana
- Assessment of e-learning readiness in National University of Mongolia
  Otgontsetseg Sukhbaatar, Tsolmon Zundui, Lodoiravsal Choimaa
- Extracting Political Networks of the Sudan from Online Newspapers
  Musa Ibrahim M. Ishag, Ho Sun Shon, Keun Ho Ryu

Three-dimensional image processing using Integral Imaging

Ganbat Baasantseren
School of Engineering and Applied Sciences, National University of Mongolia
ganbat@num.edu.mn

Abstract

Our world is three-dimensional (3-D), so researchers are developing 3-D technology to use in everyday life. In 3-D computer graphics, 3-D model is developing a mathematical representation of any 3-D surface of the object with specialized software. 3-D rendering what coverts 3-D models into 2-D images because all most all displays are two-dimensional (2-D). However, 3-D technologies are depleovping and used a narrow array such as entertainment, medical application, ingreenig design. Among 3-D technologies, an Integral Imaging (InIm) is most important because because it has advantages such as pull parallax, without special glasses, mute-viewers, and color. In this presentation, we introduce two things. Both are 3-D image processing based on InIm technology. In the first, we introduce a new method to reduce a processing time what creates an Elmental image, what is sub 2-D images of 3-D objects, for the InIm display of the 3-D model. We did not check satisfying condition because the closest elemental lens from the elemental pixel is corresponding elemental lens. It can reduce one loop for matching elemental lens so new method can reduce the processing time. From the result, the proposed method is 60 times faster than conventional methods. Also, InIm technology also can create a fully parallax image. We used this unique property to generate an arbitrary view image in the second method. It is like that we can see different images at different positions if we see the real 3-D object. Our method can create 25 different images from one set Elemental image of 3D object. It is like that we see a 3-D object at 25 positions. The elemental images have been captured by experiment..

Mining Frequent Itemsets in Transactional Database by Reduction of Trees Traversal

Supatra Sahaphong[1], Gumpon Sritanratana[2]
[1]*Department of Computer Science, Faculty of Science, Ramkhamhaeng University, Thailand.*
supatra@ru.ac.thl
[2]*Department of Mathematics, Faculty of Science, Buriram Rajabhat University, Thailand.*
sgumpon@gmail.com

Abstract

This paper is aimed to develop a new algorithm to mine all frequent itemsets with minimum support threshold from a transaction database. The new mining algorithm performs database only once and without generating any candidate itemsets. The new algorithm uses itemset tree which is not only the number of tree construction is reduction but also its recursive of mining steps is reduced. The experiments in which run time and memory consumption of the propose algorithm are tested in comparison with frequent pattern growth algorithm. The experimental results demonstrate that the new algorithm provides better performance than frequent pattern growth in terms of run time and space consumption.

Keywords*: Association rule mining; Data mining; Frequent itemsets mining; Knowledge discovering*

1. Introduction

The association rule mining is an essential task of data mining. It is to decompose into two major steps. First, the generation of all the frequent itemsets which satisfy the minimal support threshold or minsup. Second, the extraction of all high confidence rules from frequent itemsets found in previous step. Our work focuses on the first step. The first classic algorithm is Apriori which is proposed in [1]. The Apriori principle is "If an itemsets is frequent, then all of its subsets must also be frequent" [2]. The Apriori algorithm uses a level-wise and breadth-first search approach for generating association rule. It uses the support-based pruning to control the exponential growth of candidate itemsets. The algorithms based on generated and tested candidate itemsets have two major problems which are shown as follows. The database must be scanned multiple times to generate candidate sets which increase the I/O load and is time-consuming. Moreover, the generation of huge candidate sets and calculation of their support will consume a lot of CPU time. The drawbacks which presented as above were overcome by using the next generation of algorithm, called the FP-growth algorithm [3]. The advantages of mining of frequent itemsets by using the FP-growth algorithm are shown as follows. The database is scanned only two times, so time consuming is decreased. The generating of candidate sets is not required, so the I/O load is reduced. The FP-growth algorithm performs depth-first search approach in the search space. It encodes the data set using a compact data structure called FP-tree and extracts frequent pattern directly from this prefix tree [4]. The following researches have improved this idea. In reference [5], the H-mine algorithm was introduced by using array-based and trie-based data structure. The Patricia Mine algorithm was proposed in [6] that compressed Patricia trie to store the data sets. The FPgrowt* algorithm reduced the FP-tree traversal time by using array technique [7]. In reference [8], the SFI-Mine algorithm which constructs pattern-base by using a new method which is different from pattern-base in FP-growth and mines frequent itemsets with a new combination method without recursive construction of conditional FP-tree. However, most of the FP-tree algorithm base has the following drawbacks. First, mining of frequent itemset from the FP-tree, it generates huge of conditional FP-tree and takes a lot of time and space. Second, when the changing of minimum support, this algorithm may restart and scan database twice. Many researchers have proposed ways to scan database once. The Eclat algorithm was proposed by using the join step from the Apriori property to generate frequent pattern [9]. In Reference [10], the new data structure, called LIB-

graph is proposed to contain data when database is scanned and discovery of frequent patterns by using recursive conditional FP-tree. The Sorted-List structure which created from the Vertical Index List was proposed to contained data from scanning database once and mining of frequent itemsets by using depth-first search [11]. Moreover, in case that the decision maker wants to change the minimum support threshold, an algorithm is performed without rescanning of database [12].

This paper proposed a new algorithm to mine all frequent itemsets. The feature of the proposed algorithm presented as follows. The database is scanned only one time to mine frequent itemsets and a new algorithm mines frequent itemsets without generation of candidate sets. The decision maker can change of the minimum support threshold all time without rescanning of the database. The proposed algorithm reduced the number of sub-trees and loops in mining steps. Therefore, the proposed method can reduce both of run time and space consumption, the experiments in which the run time and memory consumption are test for the VIL-tree and FP-growth algorithm. The results of this method are still obtaining complete and correct frequent itemset. This paper is organized as follows. The prior knowledge is presented in section II, followed by the approach which is presented in section III, the results and discussions is shown in section V and the finally, the conclusion is addressed in section VI.

2. Prior knowledge

The basic concepts of mining frequent itemsets are presented as following.

Let $I = \{x_1, x_2, ..., x_m\}$ be a set of items and $DB = \{T_1, T_2, ..., T_n\}$ be a transaction database, where $T_1, T_2, ..., T_n$ are transactions that contain items in *I*. The support, or *supp* (occurrence frequency), of a pattern *A*, where *A* is a set of items, is the number of transactions containing *A* in DB. A pattern *A* is frequent if *A*'s support is no less than a predefined minimum support threshold, *minsup*. Given a DB and a minimum support threshold *minsup*, the problem of finding a complete set of frequent itemsets is called the frequent-itemsets mining problem. All above terminologies are proposed by Han et al [4, 12].

A data structure called a vertical index list (VIL) which introduced in [2, 9, 11] is summarized as follows. Let $T_i = \{x_1, x_2, ..., x_m\}$ be a transaction in DB, where $i = 1,2,...,m$ and x_j is an item for $j = 1,2,...,n$. A vertical index list (or VIL) is the structure constructed from a scan of each T_i in DB only once. Each row in VIL contains an item in I, support of item in I, and transactions in DB which contain such an item. The set of transaction will be written in order according to the ascending of its identification number. The set of items will be written in order according to the descending of its support. The algorithm of VIL construction is shown in [13].

3. The approach

The transaction database is scanned once to construct a VIL. Then an itemset-tree structure is a general tree structure constructed from the VIL. It is a finite set of one or more nodes. It consists of the root of tree, a set of item subtrees as the children of the root, and a set of header tables.

```
Algorithm: Mining frequent itemsets
Begin
  For each frequent item in vertical index list do
    Find a set of frequent itemset length-1 which
      supp(item) not < minsup and save all of it to
      answer-buffer-length-1
  End //For
  For all frequent itemset length-1 do
    Create itemset-tree and then create a set of
      sub-header-length-1
    Generate candidate root of sub-tree(length-1)
    For all candidate root do
      While (candidate root is not in answer-buffer-length-m
          and level of tree is  greater than 2) do
          //where m=2,3,…,n
        Create sub-tree(length-m) and then create sub-
          header-length-m
        Generate candidate root of sub-tree(lenfth-m)
      End //While
      Combine root node&child node and save it to answer-
        buffer-length-m
    End //For
  End //For
  Union all answer-buffer-length-m and save to answer set of
      frequent itemsets
  End //Begin
```

Figure 1. Mining frequent itemsets

Each node in tree comprises five fields. There are two fields of value which are item-name and support and there are three fields of pointer which are same-item, parent, and child. Each member of the header table consists of two fields, item-name and head of node link. Each node of tree is of the form (frequent itemset :support). The algorithm of itemset tree construction is presented in [14]. The proposed algorithm in Figure 1 shows how to mine all frequent itemsets.

4. Results and Discussion

This section presents the experiments in which the run time and memory consumption are test for the new algorithm and FP-growth algorithm with a synthetic datasets and varying minimum support thresholds. The experiments were performed on a Microsoft Windows 7 Home Premium, processor is (Intel (R) Core (TM) i5-2467M, and 4 GB of RAM. All algorithms were coded using C language. The synthetic datasets generated by the IBM Almaden Quest research group [15-16] were used for presented the experimental results. The datasets serve as the FIMI repository, which is a result of the workshops on frequent itemset mining implementations [16-17]. The original database of synthetic datasets is T10I4D100K.

In Figure 2, when the minimum support is high, the number of frequent itemsets is low. The minimum support is low, many frequent itemsets are obtained. A new method is always faster than FP-growth method because frequent item in vertical index list is performed in order of high support. The node construction and sub-trees are reduced, resulting in a reduction in run time and memory consumption. The Figure 3 shows that the memory consumption of a new method is less than FP-growth in every minimal support threshold.

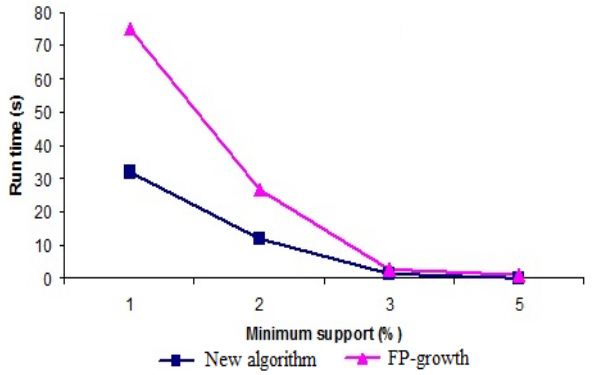


Figure. 2. Run time

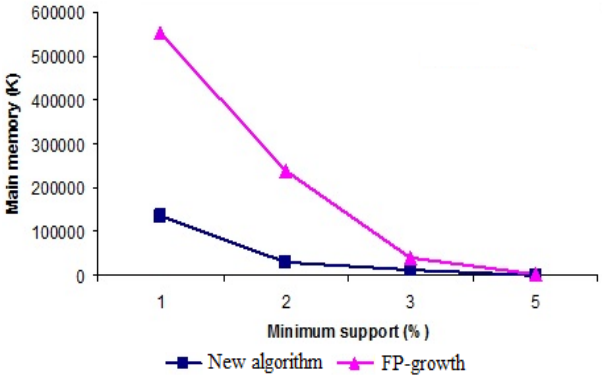


Figure 3. Memory consumption

5. Conclusion

We have developed a new algorithm to mine all frequent itemsets with minimum support threshold from a transaction database. This new mining algorithm performs database only once and without generating any candidate itemsets. The research provided the experiments in which run time and memory consumption are tested in comparison with FP-growth algorithm. The experiments of both algorithms are evaluated by applying to the bench mark synthetics datasets. We summarize the feature of this research as follows. This news algorithm scans database only once. Moreover, this algorithm uses sorted vertical index list, so amount of the frequent itemsets are generated at first. The next tree is resized down and sub-trees are reduced which reduced the number of loops of mining steps. Therefor, run time and space consumption is reduced. The experimental results demonstrate that the new algorithm provides better performance than FP-growth in terms of run time and space consumption.

Acknowledgement

We deep appreciation and gratitude to Associate Professor Dr. Veera Boonjing of the Department of Mathematics and Computer Science, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Thailand for his guidance and suggestions. We would like to express our appreciation and gratitude to Associate Professor Dr.Tawesak Tanwandee of Mahidol University, Thailand for his help in proof reading of this paper.

6. References

[1] R. Agrawal and R. Srikant: Fast Algorithm for Mining Association Rules, *Proceedings of the 20th International Conference on Very Large Databases*, Chile, September 1994, 487-499.

[2] P-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining (Pearson Education Inc., 2006).

[3] J. Han, J. Pei, and Y. Yin: Mining Frequent Pattern without Candidate Generation, *Proceedings of the 2000 ACM SIGMOD international conference on Management of Data*, Texas, May 2000, 1-12.

[4] J. Han, J. Pei, Y. Yin, and R. Mao: Mining Frequent Pattern without Candidate Generation: a Frequent Pattern Tree, *Springer*, vol. 8, 2004, no 1, 53-87.

[5] J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, and D. Yang: Hmine: Hyper-Structure Mining of Frequent Patterns in Large Databases, *Proceedings of the 2001 IEEE International Conference on Data Mining*, USA, November 2001, 441-448.

[6] A. Pietracaprina, and D.Zandolin: Mining Frequent Itemsets Using Patricia Tries, *Proceedings of the 3rd IEEE International Conference on Data Mining*, Florida, USA, November 2003.

[7] G. Grahne, and J. Zhu: Efficiently Using Prefix-Trees in Mining Frequent Itemsets," *Proceedings of the 3rd IEEE International Conference on Data Mining*, Florida, USA, November 2003.

[8] S. Sahaphong, and V. Boonjing: The Combination Approach to Frequent Itemsets Mining, *Proceedings of the 2008 International Conference on Convergence and hybrid Information Technology*, Korea, November 2008, 565-570.

[9] M.J. Zaki, Scalable Algorithms for Association Mining, *IEEE Transaction on Knowledge and Data Engineering*, vol. 12, no. 3, 2000, 372-390.

[10] D. J. Chai, L. Jin, B. Hwang, and K. H. Ryu: Frequent Pattern Mining Using Bipartite Graph, *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, Germany, August 2007, 182-186.

[11] S. Sahaphong, Frequent Itemsets Mining Using Vertical Index List, *Proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology*, China, August 2009, 480-484.

[12] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, Elsevier, Maryland Heights MO, 2006.

[13] S. Sahaphong and G. Sritanratana, Proceedings of the international Conference on Cricuits, *Systems,Signal Processing, communications and Computers*, Italy, March 2014, 213-217.

[14] S. Sahaphong and V. Boonjing, IIS-Mine: A new efficient method for mining frequent itemsets, *MIJST* 2012, 6(01), 130-151.

[15] Frequent Itemset Mining Dataset Repository, "T10I4D100K". Available: http://fimi.cs.helsinki.fi/data/

[16] Workshop on Frequent Itemset Mining Implementations. Available: http://fimi.ua.ac.be/fimi03/

[17] Workshop on Frequent Itemset Mining Implementations. Available: http://fimi.ua.ac.be/fimi04/

Assessment of e-learning readiness in National University of Mongolia

Otgontsetseg Sukhbaatar[1], Tsolmon Zundui[2], Lodoiravsal Choimaa[1]
Department of Electronics and Communication Engineering,
School of Engineering and Applied Sciences, National University of Mongolia
{otgontsetseg, lodoiravsal}@num.edu.mn
Department of Information and Computer Science,
School of Engineering and Applied Sciences, National University of Mongolia
tsolmonz@num.edu.mn

Abstract

Recent years, e-learning became a common delivery media for education all over the world. While e-learning technology introduced relatively late in Mongolian universities, there isn't any successful implementation history yet. Evaluation of the institutional soundness of e-learning as well as the readiness of learners is crucial to engage e-learning successfully. The purpose of this study was to support the development of an instrument to measure readiness in online learning environments, assess students' readiness in National University of Mongolia (NUM) and identify influencing factors among instrumental items. In this study, we used 15-item instrument with 5-point Likert-type scale answers. Results were based on responses from 400 undergraduate students covering 13 departments of NUM. The study has found that majority of the students are ready for e-learning, demonstrating high score level of preparation and five major factors were identified that highly affects students readiness.

1. Introduction

Recent years, e-learning became common delivery media for education and training and designing and implementing web-based education systems and platforms have grown dramatically. In United States of America, for example, more than 1.9 million students took at least one online course and one-third of them choose all their courses online in the Fall semester of 2003 [6]. This number has increased dramatically and surpassed 7.1 million in 2013 [7]. Herewith many universities started to develop Massive Open Online Courses (MOOC) since 2012 [7].

E-learning improves students' learning method, time management skill and promotes ability to work independently. Moreover, advantages like cost reduction, elimination of time and space constraint make it more important and popular. However, integrating online courses into curriculum cannot be done overnight, just solving technological issues. Today's Mongolian students have experience in traditional classroom learning, but may not have experience in online learning environments. Therefore, it is essential to evaluate e-learning system's different aspects and understand factors which influence its effectiveness. In this work, the following three main questions are considered:

- How much students ready for e-learning in NUM?
- What can be the important factors affecting students' readiness greatly?
- What should we concern to ensure successful implementation of e-learning?

2. Background

An online course is defined as one in which at least 80 percent of the course content is delivered online [7]. Few years back, it's assumed that online courses are available only in higher education institutions. Today, many universities offer online undergraduate and graduate courses for students and MOOCs for those outside of the institution's student body. Since effective use of technology in delivering the curriculum has begun to take importance in many universities, National University of Mongolia did not pass over the movement toward e-learning. From 2012, School of Engineering and Applied Sciences allowed students to collect credits taking Small Private Online Courses (SPOCs) in edX, which is MOOC platform founded by MIT and Harvard University. It clearly

showed that student's learning style, self-sustainability affects online learning effectiveness considerably. And there were less issue in terms of technical access and use of technological tools, because students were from engineering school.

Readiness for e-learning refers to three major aspects, (a) students' preferences for online learning as opposed to face-to-face learning instructions, (b) students' capability and confidence in using technological tools, (c) students' ability to learn independently [8]. In order to assess the e-learning readiness, it is crucial to have a valid and internally consistent questionnaire as instrument. Although, the development of instrument items on literature review is necessary, it is pretty important to define the factors of the effective measure and determine internal consistency of the recommended items.

Self-assessment is one way of gauging potential online-learner's readiness. R. Watkins, D. Leigh and D. Triner used 27 statements (6 categories) with 5-point Likert-type scale answers [1]. In a study at the Business School in a Malaysian Private University, authors used 4-point 13 items, which can be grouped in 4 main categories [2]. In this work, we developed 15 questions based on the works mentioned above, adjusting to the circumstances of Mongolia. Questions grouped into 4 categories and have a 5-point Likert-type scale responses.

3. Research methodology

In this study, survey strategy is adopted. The survey is popular and common strategy in business research that is usually associated with deductive approach [9]. Survey allows the collection of larga amount of data from a sizeable population in a highly economical way. Random sampling without replacement has been used in this study because the main population is not so large, requires minimum advance knowledge of the population other than frame, and it is free of classification error.

The sample of this study was made up from 400 students from National University of Mongolia (NUM). After reviewing the collected paper based questionnaire data, it was found that 399 students were filled completed and one student was missing one answer. Therefore, 399 responses were entered into a data file and analyzed with SPSS.

Questionnaire

The questionnaire has been designed to measure students' socio-demographic characteristics including age, gender, enrollment date, school, as well as the four factors including self-study management, reflective thinking, interaction support and learning setting. In addition, the questionnaire comprised 19 items distributed over two sections. Part II included factors to be evaluated using a five-point scale, ranging from strongly disagree (1), disagree (2) neutral (3), agree (4), and strongly agree (5).

The data from students of this sample were obtained by asking them to fill a paper based questionnaire. To help the respondents truly understand everything that was being asked, designed to be as easy and was in Mongolian. Also it was not too long, so it could be filled out within 10 minutes. Questionnaire used in this study is shown in Table 1.

Table 1. Questionnaire

Group		Items	Mean
Technology Access	a1	I have a personal computer with an Internet connection at home.	4.21
	a2	My computer is fairly new (performance is good enough).	3.99
	a3	My computer has an adequate softwares for e-learning environment (browser, Microsoft Word, Adobe Reader, etc).	4.24
	a4	I have a basic skills for finding my way around the Internet (using search engines, browsers, login using accounts, etc).	4.29
	a5	I think that I would be able to communicate effectively with others using online technologies (email, chat, forum, discussion, etc).	4.49
Interaction Support	b1	I think that online learning is at least equal quality to traditional classroom learning.	3.52
	b2	I am willing to actively communicate electronically with my classmates and instructors who are in different time zones.	3.93
Reflective thinking	c1	I feel that my background and experience will be beneficial to my studies.	4.49
	c2	I am comfortable with written communication.	3.42
	c3	I believe looking back on what I have learned in a course will help me to understand it better.	4.47
Self-study management	d1	I think that I would be able to complete my assignments on time even when there are distractions (television, chat with friends, web surfing, etc).	4.06
	d2	I am able to manage my study time effectively and complete assignments on time without someone's instructions	4.11
	d3	I usually set goals and have a high degree of initiative for my studies.	4.13
	d4	I am willing to dedicate 8-10 hours per week for online learning.	3.89
	d5	I enjoy working/studying independently.	4.17

4. Data analysis

The data collected was analyzed using the IBM SPSS version 19. Descriptive statistics were used as a mean of describing the socio-demographic characteristics of the students, as well as their pe9rceptions of the variables.

4.1. Socio-Demographic characteristics

Figure 1 represents the distribution of students across schools of NUM. The study reveal that the majority of the respondents were not IT related majoring students. Only 11 percent respondent students were majoring IT.

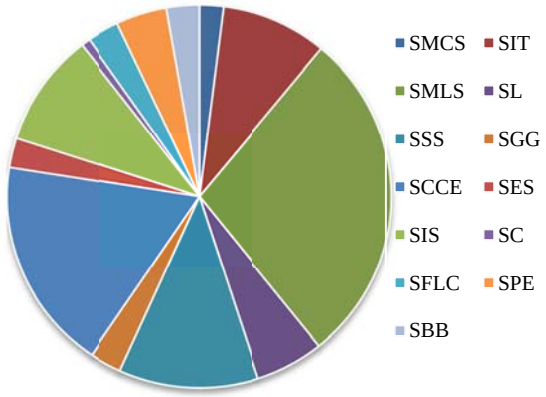


Figure 1. Fields of study preferences

Figure 2 represents the distribution of the students by school year. As shown, most of the sample consisted of first year students.

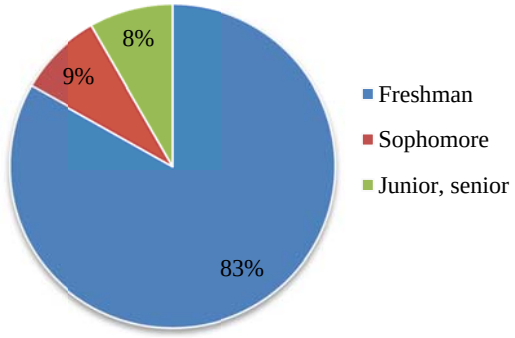


Figure 1. Year distribution of students

Figure 3 represents the distribution of respondents across age. It reveals that 46 percent of the respondents were age of 18, with the minimum age being 15 and the maximum of 21 years old.

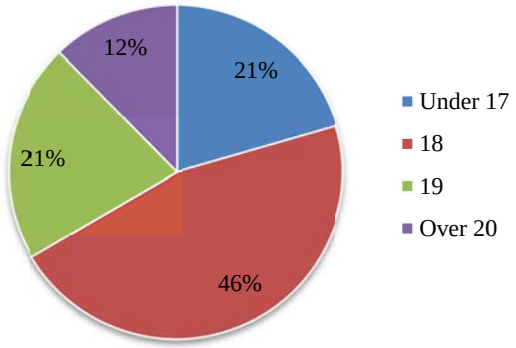


Figure 2. Age spectrum of students

4.2. Reliability analysis

Cronbanch's Alpha reliability test was performed on 15 items, and the obtained results are presented in Table2. Value, which are greater than 0.7, were taken as indicators of reliability for these measures. The commonly accepted threshold value for social science is that alpha should be 0.70 or higher for a set of questions to be considered reliable because at alpha 0.70, the standard error of measurement will be over half of a standard deviation. When tested, the overall questionnaire showed a high value of Cronbanch's Alpha reliability coefficient as shown in Table 2.

Table 2. Reliability characteristics

Cronbach's Alpha	N of Items
.810	15

4.3. Correlation analysis

The value for a Pearson's correlation can fall between 0.00 (no correlation between variables) and 1.00 (direct correlation between variables). Knowing the value on one of the variables provides no assistance in predicting the value on the second variables.

As shown in Table 3, correlation table has no correlations that are >0.70, which shows that multi-collinearity is not a significant program in this case. Therefore, there is no need to consider eliminating any variables.

Table 3. Correlation table

	a1	a2	a3	a4	a5	b1	b2	c1	c2	c3	d1	d2	d3	d4	d5
a1	1.00	.593	.507	.263	.293	.201	.172	.059	.066	.050	.145	.127	.116	.309	.149
a2	.593	1.00	.593	.314	.281	.193	.204	.075	.062	.021	.156	.170	.111	.252	.183
a3	.507	.593	1.00	.374	.340	.210	.185	.200	.076	.103	.184	.217	.158	.285	.160
a4	.263	.314	.374	1.00	.440	.122	.179	.144	.019	.069	.141	.175	.197	.223	.119
a5	.293	.281	.340	.440	1.00	.211	.142	.289	.012	.144	.186	.244	.249	.291	.147
b1	.201	.193	.210	.122	.211	1.00	.376	.155	.095	.077	.272	.239	.183	.364	.217
b2	.172	.204	.185	.179	.142	.376	1.00	.272	.050	.182	.279	.231	.202	.234	.288
c1	.059	.075	.200	.144	.289	.155	.272	1.00	.193	.317	.341	.288	.403	.226	.348
c2	.066	.062	.076	.019	.012	.095	.050	.193	1.00	.255	.134	.191	.191	.135	.249
c3	.050	.021	.103	.069	.144	.077	.182	.317	.255	1.00	.231	.193	.240	.226	.262
d1	.145	.156	.184	.141	.186	.272	.279	.341	.134	.231	1.00	.495	.523	.357	.368
d2	.127	.170	.217	.175	.244	.239	.231	.288	.191	.193	.495	1.00	.495	.297	.417
d3	.116	.111	.158	.197	.249	.183	.202	.403	.191	.240	.523	.495	1.00	.322	.464
d4	.309	.252	.285	.223	.291	.364	.234	.226	.135	.226	.357	.297	.322	1.00	.394
d5	.149	.183	.160	.119	.147	.217	.288	.348	.249	.262	.368	.417	.464	.394	1.00

a. Determinant = ,019

4.4. Factor analysis

The researchers [2][3][5] did not have strong theory about the constructs underlying responses to their measures. Therefore, we did an exploratory factor analysis.

Table 4. Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	Variance % of	Cumulative %	Total	Variance % of	Cumulative %	Total	Variance % of	Cumulative %
1	4.329	28.862	28.862	4.329	28.862	28.862	2.677	17.850	17.850
2	1.983	13.223	42.085	1.983	13.223	42.085	2.574	17.157	35.007
3	1.105	7.364	49.449	1.105	7.364	49.449	1.865	12.430	47.437
4	1.066	7.106	56.555	1.066	7.106	56.555	1.368	9.118	56.555
5	.960	6.400	62.955						
6	.795	5.301	68.256						
7	.746	4.974	73.231						
8	.667	4.446	77.677						
9	.653	4.352	82.029						
10	.547	3.648	85.677						
11	.514	3.425	89.103						
12	.471	3.142	92.244						
13	.416	2.772	95.016						
14	.395	2.636	97.652						
15	.352	2.348	100.000						

Extraction Method: Principal Component Analysis.

Table 4 lists the eigenvalues associated with each linear factor before extraction, after extraction and after rotation. Before extraction, 15 linear components were identified within the data set. The eigenvalues associated with each factor represent the variance explained by the that particular linear component and percentage of variance explained, so factor 1 explains 28.862 % of total variance.

In this, 15 factors would be needed to explain 100% variance in the data. Four components were extracted, then 56.666% of the variance would be explained.

Table 5 provides a far more interpretable solution that varimax rotation because the difference between high and low loading is more apparent in the pattern matrix, which eliminates the complex variables and has a simpler structure. There are several things to consider about the format of this matrix. First, factor loading less than 0.4 have not been displayed because I asked for these loadings to be suppressed. Second, the variables are listed in the order of the size of their factors loading because I asked for the output to be sorted by size.

Table 5. Rotated Component Matrix

	Component			
	1	2	3	4
d3	.726			
c1	.647			
d2	.599			
d1	.583			
a5	.549			
d5	.459			
a2		.808		
a3		.785		
a1		.776		
a4		.552		
b1			.762	
b2			.655	
d4			.472	
c2				.796
c3				.533

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 10 iterations.

4.5. Multiple regression analysis

The main purpose this study was to determine what factors affects e-learning readiness of students at NUM. To achieve this goal, a regression analysis is used. In this study, simple linear regression is the most appropriate means to compare a set of factors.

The regression model is represented by the following equation:

$$Y_i = B_0 + \sum_i^p B_i * X_{ij} + \varepsilon_i \quad (1)$$

where, for i^{th} case, Y is the response variable, X_{ij} are p repressor, and u_i is a mean zero error. The quantities b_is are unknown coefficients.

Table 6. Regression coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Cons)	-.182	.343		-.530	.596	-.857	.493
a1	.173	.037	.202	4.702	.000	.101	.246
b1	.227	.045	.223	5.016	.000	.138	.316
c3	.141	.061	.102	2.320	.021	.021	.260
d1	.193	.057	.158	3.395	.001	.081	.304
d5	.272	.055	.231	4.973	.000	.164	.379

Based on the result shown in Table 6, the main factors determined by this study are a1, b1, c3, d1 and d5. The other 10 factors can be removed. The analysis shows the importance of all factors have an effect. The e-learning readiness of the students is represented by the equation:

$$readiness = -0.182 + 0.173a_1 + 0.227b_1 + 0.141c_3 + 0.193d_1 + 0.272d_5 + \varepsilon \quad (2)$$

Table 7. Main factors

Group		Items	Mean
Technology Access	a1	I have a personal computer with an Internet connection at home.	4.21
Interaction support	b1	I think that online learning is at least equal quality to traditional classroom learning.	3.52
Reflective thinking	c3	I believe looking back on what I have learned in a course will help me to understand it better.	4.47
Self-study management	d1	I think that I would be able to complete my assignments on time even when there are distractions (television, chat with friends, web surfing, etc).	4.06
	d5	I enjoy working/studying independently.	4.17

5. Conclusion

The study found that NUM students are ready for e-learning with 77 percent of readiness.

We identified five main factors that affect e-learning readiness of students greatly, as follows:

1. To have a personal computer with an Internet connection at home
2. Online course's good content and quality compared to classroom learning
3. Possibility of watching course materials repeatedly
4. Students' self-management skill
5. Students' tendency for working independently

Table 7 shows these 5 factors with mean values. While comparing mean values of the factors, c3 had minimal value of 3.52.

Result of this study shows that it is crucial to consider students' independent learning, self-management and online communication skills rather than technical requirements. Majority of the students gave high point responses to questions in Technology Access group.

In order to successfully implement e-learning, we should provide knowledge of its advantages compared with classroom learning and develop students' skills of independent studying.

Further, we will conduct research on institutional readiness in terms of organization and faculty members to complete the study.

6. References

[1] R. Watkins, D. Leigh andD. Triner, "Assessing readiness for e-learning" *Performace Improvement Quarterly*, vol. 17, No.4, 2004 , pp. 66-79.

[2] S. F. Tang and C. L. Lim, "Undergraduate students' readiness in e-learning: A study at the business school in a Malaysian Private University", *International Journal of Management and Information Technology*, vol. 4, No.2, July 2013, pp.198-204.

[3] A. Keramati, M. Afshari-Mofrad and A. Kamrani, "The role of readiness factors in e-learning outcomes: An empirical study", *Computers& Education 57*, 2011, pp. 1919-1929.

[4] B. Darab and G.A. Montazer, "An eclectic model for assessing e-learning readiness in the Iranian universities", *Computers &Education 56*, 2011, pp. 900-910.

[5] J. Mahat, A. F. Mohd Ayub, S. Luan, Wong,"An assessment of students' mobile self-efficacy, readiness and personal innovativeness towards mobile learning in higher education in Malaysia", *International educational technology conference (IETC2012)*, Procedia-Social and Behavioral Sciences 64, 2012, pp.284-290.

[6] E. Allen, J. Seaman, "Sizing the opportunity: The quality and extent of online education in the united

states, 2002 and 2003", *The Sloan Consortium*, September 2003, from http://sloanconsortium.org/publications/survey/sizing_the_opportunity2003.

[7] E. Allen, J. Seaman,"Grade change: Tracking online education in the united states", *The Sloan Consortium*, 2013, from http://www.onlinelearningsurvey.com/reports/gradechange.pdf

[8] D. Warner, G. Christine, S. Choy, "The readiness of the VET sector for flexible delivery including on-line learning", *Brisbane: Australia National Training Authority*.

[9] Groves, R.M.; Fowler, F. J.; Couper, M.P. (2009). *Survey Methodology*. New Jersey: John Wiley & Sons.

[10] Cronbach L.J (1951). "Cefficient alpha and the internal structure of tests". *Psychometrika* 16 (3): 297-334.

[11] Francis, DP; Coast AJ, Gibson D (1999). "How high can a correlation coefficient be?". *Int J Cardiol* 69 (2): 185-199.

Extracting Political Networks of the Sudan from Online Newspapers

Musa Ibrahim M. Ishag, Ho Sun Shon, Keun Ho Ryu
Database/Bioinformatics Laboratory, Department of Computer Science,
Chungbuk National University, South Korea
{ibrahim, shon0621, khryu}@dblab.chungbuk.ac.kr

Abstract

The rapid proliferation in the intersection between data mining, information retrieval, and social network analysis has made it possible to analyze unstructured big-data at a massive scale. In this paper, motivated by the sheer amount of unstructured big data generated by online newspapers, we are proposing a framework for extracting political networks of the Sudan from news articles. In essence, our proposal consists of a crawler that collects news articles from famous online news outlets, a preprocessing module and a network formulation module. The crawler is able to fetch the news as it is updated. Therefore, it feeds the system with useful data. The preprocessing module includes named entity recognition for identifying the key players and an event extraction to detect political events. In addition, the preprocessing module is able to resolve duplicates. The network module initiates the actual network of the names and events detected. The framework resembles the base for future structural and content analysis of political networks.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2008-0062611). And The International Science and Business Belt Program through the Ministry of Science, ICT and Future Planning (2013K001552). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NRF, Korean Ministry of Science, or the Korean Government.

Session : Data Mining Applications

- Fuzzy Measure Application to Decision Making
  Sanghyuk Lee
- Real-Time Data Warehousing and Online Analytical mining of redesigned large database: Challenges and Solutions
  Oyun-Erdene Namsrai
- Differential Wheeled Mobile Robot Self-localization Method for 8 Bit Microcontroller
  Batbayar Unursaikhan, O.Zoljargal

Fuzzy Measure Application to Decision Making

Sanghyuk Lee
Dept. of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, China
Sanghyuk.Lee@xjtlu.edu.cn

Abstract

General way of fuzzy measure and integral were introduced. Relation with meta-heuristic measure and issues such as decision making and classification was also investigated.

1. Introduction

Fuzzy measure and integral application to decision making has been studied by numerous researches [1, 4 and references there in]. From the introduction of Sugeno, fuzzy measure and integral has provided an important background for theoretical and practical point of view[2]. Actually, monotonic non-additive set function was considered as fuzzy measure by Sugeno, and it has been appeared as achievements of Choquet's and in cooperative game theory [3, 4].

In order to get reasonable decision value, it is required to conduct the calculation based on optimization or rational approach. There are several ways of conducting optimization such as least mean square algorithm. With the first application of the concept of fuzzy measure to the field of multi-criteria decision making [2], fuzzy measure has become an efficient and validity method of evaluation real problem. By considering the fuzzy set A, in which elements have various membership values. For fuzzy set A,

$A = \{x | x \in X, 0 \leq \mu_A(x) \leq 1\}$, where X is universe of discourse.

It means that fuzzy set can provide more general expression including individual opinion [5]. By calculating of the fuzziness, we could evaluate degree of uncertainty and similarity through designing of fuzzy entropy and similarity measure [6-10]. With the help of fuzzy entropy and similarity, we can also provide solution of clustering, reliable data selection. Because of these flexible characteristics, it constituted the expert system with neural network. Specially, it provided very efficient supporting tool for human decision of strategies.

Decision theory has been studied as rationale basis; its related topics deal optimization with/without constraint, linear and nonlinear optimization. Including control strategies, multiple strategies can be composed as candidates of solution. In order to get solution, it is also needed to consider constraint, whether it is uni/multi-criteria. Furthermore, is it coalition or not? We studied decision under the condition of multi-criteria, which means that we have to consider multiple constraint or condition to get proper decision. By analyzing an illustrative example, a complete system of making decision will be established to the reader. This example acts as the main clue of the report and the author will establish all relative values with detail process of calculation.

Shapley index and Interaction index are calculated to show the process of making decision and then the model outputs are calculated. However, several problems are coming with the output of the model; they are expected to be solved with further study.

2. Preliminaries

Some important terminologies and methods are briefly introduced.

2.1. HLMS

HLMS is short for heuristic least mean square, which is a new algorithm raised by Michel Grabisch [5]. It solves the problem that a 2^n representation is difficult to calculate for a large number of 'n' and the standard optimization algorithms are not suitable to fuzzy measures. It also has a superior behavior than the algorithm raised by Mori and Murofushi [6].

This algorithm uses the obtained data to do a series calculation and represent fuzzy measures in the finite case by using a lattice representation. Combine with the example just illustrated in the last part, the following figure is the illustration of the lattice.

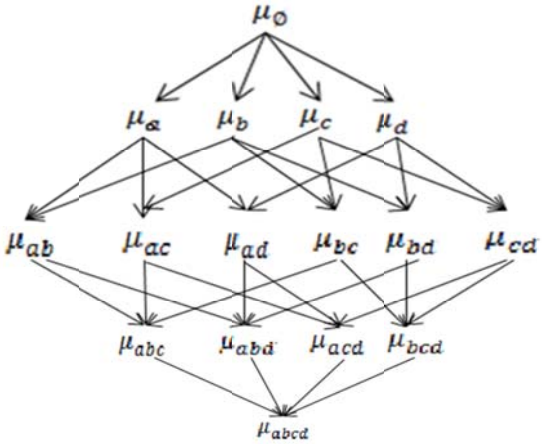


Figure 1. Lattice structure of four criteria

2.2 Multi-criteria

Multi-criteria means that several criteria affect the decision of the instructor at the same time. But the effect of one criterion is different from the other one. In the Multi-criteria case, the algorithm firstly offers the data and then the importance of every criterion and their relationship can be established in a clear form by using Shapley index and Interaction index.

The multi-criteria analysis is different from the uni-criterion method. In the second case, only one criterion affects the decisions which do not need other methods to verify whether it is the best choice. However, in the multi-criteria case, there need other method to verify it, which is caused by the fact that there are several criteria influence the choice simultaneously and we do not know the priority between those criteria.

2.3 Coalition

A coalition is a treaty among groups or individuals (between two criteria in this case), during which they have joint part [4]. It has significant effects on calculating the interaction index, which is most informative. With the instruction of the interaction index, the importance of the coalition is obviously and can help the reader to judge and evaluate the behavior of the instructor.

2.4 Shapley index

Shapley index can be considered as an average value of one item that contributed to the whole system, it is an important index that helps decision maker to understand the importance of every criterion.
It can be calculated by the following equation:

$$v_i = \sum_{k \subset x \setminus i} \frac{(n-|K|-1)!|K|!}{n!} [\mu(K \cup \{i\}) - \mu(K)] \quad (1)$$

K is a collection of all possible combinations of criterion except criterion i.

2.5 Interaction index

Interaction index is used to display the relationship between different elements, it also has the equation to compute numerical value.

$$I_{ij} = \sum_{K \subset X \setminus \{i,j\}} \frac{(n-|K|-2)!|K|!}{(n-1)!} [\mu(K \cup \{i,j\}) - \mu(K \cup \{j\}) - \mu(K \cup \{i\}) + \mu(K)] \quad (2)$$

Again, K is a collection of all possible combination of elements except i and j and $|K|$ is the number of members of set K.

3. Application to Example

In this example, we consider the problem of the selection of electrical products; different kinds of products are evaluated based on the following four criteria, cost, life span, popularity and stability. Fors cost and lifespan were given as numeric values. And popularity and stability were given as evaluation. The score going from A(Recommend) To E(Not recommend) and A(perfect match) to E(Unstable). In this example five products are considered, name from A to E. The following table displays their performance on every criterion.

Table 1. Product evaluation over 4 criteria

Name	Cost	Life span	Popularity	Stability
A	550	10	B	D
B	700	8	B	B
C	500	7	C	A
D	900	6	B	B
E	850	11	C	B

The instructor can make following judgments based on the following criteria. (Cost): Under 600 CNY is best choice and over 1000 is unacceptable. (Life span): Less than 2 years is totally unacceptable and over 10 years is the best choice. Just follow the qualitative score given by the instructor.

Based on the judgments given by the instructor, we can draw the utility curves and derive the degrees of satisfaction. The utility curves are displayed in the Figure 2.

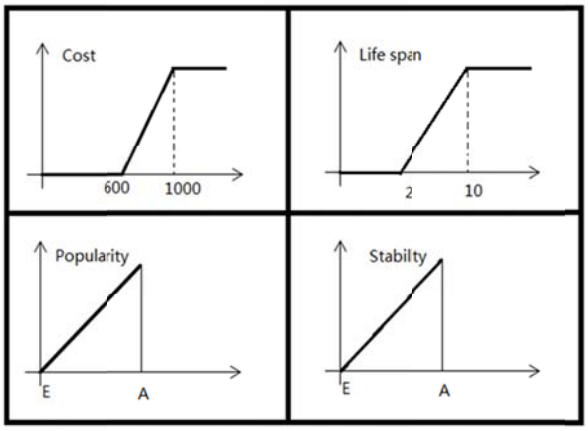


Figure 2. Utility function of 4 criteria

The numerical score based on different criteria can be given in the following table.

Table 2. Evaluation of 4 criteria via utility function and global score

Name	Cost (CNY)	Life span(Year)	Popularity	Stability	Global score
A	1.000	1.000	0.750	0.250	0.133
B	0.750	0.750	0.750	0.750	0.917
C	1.000	0.625	0.500	1.000	0.833
D	0.250	0.500	0.750	0.750	0.267
E	0.375	1.000	0.500	0.750	0.575

4. Important indexes and values

4.1 Global score

Global score is decided by the instructor. And the following paragraph will give detail description about how to divide those five products into three groups.

For the product A, it has a cheap price and long life span and there are a large number of people recommending this product, however, the degree of satisfaction to our system is only 0.25 which is unacceptable. Thus, it belongs to the last class.

Product B has an acceptable price and long enough life span, while many people also recommend this one. For the stability, it can fit our system, but there also exist the possibility that the system will break down which makes this product belongs to first class. The balance distribution of score make it is the best choice for the instructor.

Product C costs least to the instructor, but it have a relative short life span. And there are few people recommending this one which indicated some unknown latent shortcomings. This one can fit the system very well, almost no possibility to trigger the failure, which also makes it belongs to the first class. But based on overall consideration, it has less validity and practicability than product B.

Product D costs most in five products a relative short life span, but some customers also recommend this one due to its high stability. Based on the cost and life span it should be put into the last class but the high stability makes it better than A.

Product E just costs slightly less than D but it has the longest life span. For the popularity, there are several people ever used it which make the information may not be accurate but it has a relative high performance in stability. Considering four criteria, it will be label as average.

Then we can obtain the priority among five products, that is:

$$B > C > E > D > A$$

B and C belong to the first class, E is in the average level while D and A come from the third class. Because the fuzzy measures regard 1 as the fully satisfied and 0 as totally unacceptable, we can put these five points on one line distribution and then the global score can be obtained. We can put the global score interval for first class to be [1, 0.75], [0.75, 0.4] for the average level and [0.4, 0] for the third class [5]. Then the following figure will help instructor to obtain the detail global score.

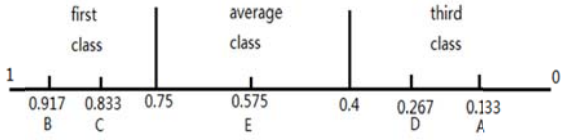


Figure. 3

4.2 Calculation process

4.2.1 Lattice representation

First we have to set $\alpha = \beta = 0.05$ and iteration to be 300. Iteration means the times that every point been modified or verified according to the HLMS. In this case, every point conducts a series of steps that just introduced for 300 times.

Table 3. values of the point

$\mu_{\emptyset}$	0
μ_a	0.16929
μ_b	0.175353
μ_c	$-9.11165e-08$
μ_d	$-4.55582e-08$
μ_{ab}	0.175396
μ_{ac}	0.175396
μ_{ad}	0.613114
μ_{bc}	0.175396

μ_{bd}	0.425354
μ_{cd}	$-4.55582e-08$
μ_{abc}	0.175396
μ_{abd}	0.787705
μ_{acd}	0.739775
μ_{bcd}	0.425375
μ_{abcd}	1

4.2.2 Shapley index

Recall the equation (1), Shapley index for i = a, x = {b, c, d}, K = ∅, {b}, {c}, {d}, {b, c}, {b, d}, {c, d}, {b, c, d} have the following eight kinds of combination, respectively.

$$\frac{(4-0-1)!\,0!}{4!}[\mu(a)-\mu(\emptyset)]=\frac{1}{4}(0.16929-0)$$

$$\frac{(4-1-1)!\,1!}{4!}[\mu(ab)-\mu(b)]=\frac{1}{12}(0.175396-0.175353)$$

$$\frac{(4-1-1)!\,1!}{4!}[\mu(ac)-\mu(c)]=\frac{1}{12}(0.175396+9.11165^{-8})$$

$$\frac{(4-1-1)!\,1!}{4!}[\mu(ad)-\mu(d)]=\frac{1}{12}(0.613114+4.55582^{-8})$$

$$\frac{(4-2-1)!\,2!}{4!}[\mu(abc)-\mu(bc)]=\frac{1}{12}(0.175396-0.175396)$$

$$\frac{(4-2-1)!\,2!}{4!}[\mu(abd)-\mu(bd)]=\frac{1}{12}(0.787705-0.425354)$$

$$\frac{(4-2-1)!\,2!}{4!}[\mu(acd)-\mu(cd)]=\frac{1}{12}(0.739775+4.55582^{-8})$$

$$\frac{(4-3-1)!\,3!}{4!}[\mu(abcd)-\mu(bcd)]=\frac{1}{4}(1-0.425375)$$

The summation of above eight equations is 0.343535. Applying this method and we can get the Shapley index for the other criteria which are displayed in the following table.

Table 4. Value of Shapley index

Shapley index	Value
Cost	0.343535
Life span	0.209463
Popularity	0.064143
Stability	0.382806

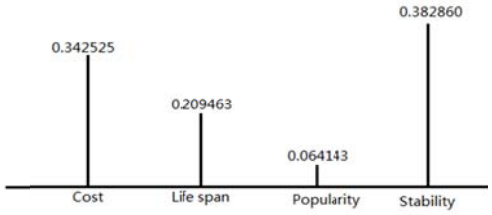


Figure 4. Distribution of data in table 6

From this figure, it is not hard to find out that stability is the most important criterion when conducting the decision and popularity has the least significance. For the cost and life span, they influence the decision made by the instructor, however, it does not as important as stability. Then we can make the conclusion that cost and life span show their importance only under the condition that the degree satisfaction of stability is large enough (0.75 and larger, indicated by the instructor)

4.2.3 Interaction index

Recall equation (2), if we set i = a and j = b then x = {c, d}, then there are following four kinds of possibility.

$K = \emptyset$;

$$\frac{(4-0-2)!\,0!}{3!}[\mu(ab)-\mu(a)-\mu(b)+\mu(\emptyset)]=\frac{1}{3}(0.175396-0.16929-0.175396+0)$$

$K = \{c\}$;

$$\frac{(4-1-2)!\,1!}{3!}[\mu(abc)-\mu(ac)-\mu(bc)+\mu(c)]=\frac{1}{6}(0.175396-0.175396-0.175396-9.11165^{-8})$$

$K = \{d\}$;

$$\frac{(4-1-2)!\,1!}{3!}[\mu(abd)-\mu(ad)-\mu(bd)+\mu(d)]=\frac{1}{6}(0.787705-0.613114-0.425354-4.55582^{-8})$$

$K = \{c, d\}$;

$$\frac{(4-2-2)!\,1!}{3!}[\mu(abcd)-\mu(acd)-\mu(bcd)+\mu(cd)]$$
$$=\frac{1}{3}(1-0.739775-0.425375-4.55582^{-8})$$

The summation of above four parts is -0.182492 Applying this method and we can get the interaction index for the other criteria which are displayed in the following table.

Table 5.Values of interaction index

Interaction index	Value
a, b	-0.182492
a, c	0.093900
b, c	0.027545
a, d	0.493931
b, d	0.239819
c, d	0.090860

From this table we can find that the interaction of 'a' and 'b' is negative while others are positive. The negative one indicates that criterion 'a' and 'b' can compensate each other. And there are significant positive interaction index between 'a', 'd', and 'b', 'd', which display the fact that stability is the most important criterion.

4.2.4 Model output

The Shapley index and interaction index influence the output of the model. There is an equation of Choquet integral for 2-additive measures, that is:

$$C_\mu(t_1 \dots \dots t_n)=\sum_{I_{ij}>0}(t_i \wedge t_j)I_{ij}+\sum_{I_{ij}<0}(t_i \vee t_j)|I_{ij}|+\sum_{i=1}^{n} t_i(v_i-0.5\sum_{j\neq i}|I_{ij}|) \quad (3)$$

Equation can be divided into three parts, the following procedures use product A as an example to display the process of calculating. The values for criteria on A are 1, 1, 0.75 and 0.25. Recall the values in table 5 and table 6, the calculation can be processed without problem. For the first two parts the result is

$$0.75\times 0.0939+0.75\times 0.027545+0.25\times 0.493931+0.25\times 0.239819+0.25\times 0.090860+1\times|-0.182492|=0.29723625$$

The last part has four conditions.

'i'=1

$$1\times(0.343535-0.5\times(0.182492+0.09390+0.493931))=-0.0416262$$

'i'=2

$$1\times(0.209463-0.5\times(0.182492+0.027545+0.239819))=-0.015465$$

'i'=3

$$0.75\times(0.064143-0.5\times(0.09390+0.027545+0.090860))=-0.031507$$

'i'=4

$$0.25\times(0.382860-0.5\times(0.493931+0.239819+0.090860))=-0.007361$$

The summation of above three items is 0.383769, and by the same procedure we can obtain the other four model outputs. The relative data are established in the following table.

Table 6. Output of the lattice representation and desire value

Product name	Model output	Desire output
A	0.383769	0.133333
B	0.750002	0.916667
C	0.830721	0.833333
D	0.368300	0.266667
E	0.567847	0.575000

4.3 Analysis

From the lattice representation we can obtain the model outputs which are given in table 6 and the order of choosing is then: C>B>E>A>D which does not satisfy the instruction given by the decision maker. There exist two inversions when comparing with the original sequence; however the two inversions are very close. This phenomenon is affected by the method of choosing global score, or the instructor's decision is not good enough. The primary goal of this algorithm is to help to verify whether the decision made by the instructor is correct or not. If there are inversions, the decision may not be the best one. There comes the problem, once the instructor changes their minds (the order of products in this case), the global score will change and the whole model should be modified. This means that we may need to build several models before the model results satisfy the instructor's willing. At the same time, the value of parameter α and β and the number of iteration will also affect the model, whose

influences are still need to be discovered with further study.

5. Conclusions

With the help of the algorithm, the problem has been put into the model; however, the model output does not match the desire output. HLMS algorithm is an estimation of the model error, thus only when the error is small enough that the model becomes reliable. In this case the error is 0.019493 which is acceptable and indicates that the output is correct to some extent. Meanwhile, this also reveals the fact that instructor's judgment is very important to the result of the model. With the modification of the global score, the model output will change as well. To make the model much more close to the real conditions, it is better for the instructor to consider the example thoroughly; after all, the model is used to verify the correctness of the decision maker, if two results cannot match, the product selected will not reach the expectation. As a conclusion, we recommend the instructor to take another evaluation method due to the poor similarity between two results.

6. References

[1] Nash, S.G. and A. Sofer, *Linear and nonlinear programming*, McGraw-Hill, 1996.

[2] M. Sugeno, *Theory of fuzzy integrals and its application*, PhD thesis, Tokyo Institute of Technology, 1974.

[3] Choquet, G. "Theory of Capacities", Annales de l'Institut Fourier, Vol. 5, pp. 131-295, 1953.

[4] Shapley, L.S. "A value for n-Person Games", in *Contributions to the Theory of Games*, Vol. II. Eds. H.W. Kuhn and A.W. Tucker. Pp. 307-317, Princeton University Press, 1953.

[5] M. Grabisch, A new algorithm for identifying fuzzy measures and its application to pattern to pattern recognition. In *Int. Joint Conf. of the 4th IEEE Int. Conf. on fuzzy systems and 2nd Int. Fuzzy Engineering Symposium*, pages 145-150, Yokohama, Japan, March 1995.

[6] T. Mori, T. Murofushi, an analysis of evaluation model using fuzzy measures and the Choquet integral, *5th Fuzzy System Symposium*, Kobe, Japan, June 2-3 1989 (in Japanese).

[7] L. Xuecheng, Entropy, distance measure and similarity measure of fuzzy sets and their relations, *Fuzzy Sets and Systems*, vol. 52, 1992, pp. 305-318

[8] D. Bhandari and N.R. Pal, Some new information measure of fuzzy sets, *Inform. Sci*. vol. 67, 1993, pp. 209–228

[9] A. Ghosh, Use of fuzziness measure in layered networks for object extraction: a generalization, *Fuzzy Sets and Systems*, vol. 72, 1995, 331–348

[10] S.H. Lee, W. Pedrycz, and Gyoyong Sohn, Design of Similarity and Dissimilarity Measures for Fuzzy Sets on the Basis of Distance Measure, *International Journal of Fuzzy Systems*, vol. 11, 2009, pp. 67-72.

[11] S.H. Lee, K.H. Ryu, G.Y. Sohn, Study on Entropy and Similarity Measure for Fuzzy Set, *IEICE Trans. Inf. & Syst.*, vol. E92-D, Sep. 2009, pp. 1783-1786.

Real-Time Data Warehousing and Online Analytical Mining of Re-designed Large Database: Challenges and Solutions

Oyun-Erdene Namsrai
School of Engineering and Applied Sciences, National University of Mongolia, Ulaanbaatar, Mongolia
oyunerdene@num.edu.mn

Abstract

Nowadays data volumes are exponentially growing and accurate business intelligence is also constantly increasing. Batches for data warehouse needs to be loaded as fresh as possible. Traditional approaches are becoming harder to solve these problems. Solutions for Real–Time Data Warehousing and Developing Decision support systems for these real time data is time consuming and expensive specially for Very Large Databases. We observed not the whole content of Database needs to be analyzed in real time. So we offered re-designed database and some of these data flows in real-time for further analysis. Also this research aims at discussing the various aspects of mining of data stored in real-time data warehouses. Various issues like data summarization, multi-objective mining and multi-level mining are also focused upon. At the end, the challenges involved in this domain are also highlighted. Finally this work summarizes our research and development activities within last year.

Differential wheeled mobile robot self-localization method for 8 bit microcontroller

U.Batbayar, O.Zoljargal
National University of Mongolia,
School of Engineering and Apply Sciences
Department of Electronics and Communication Engineering
batbayar.unursaikhan@gmail.com

Abstract

Purpose of this paper is simplifying self localization algorithm of differential driving mobile robot in order to make robot's function smarter. Rotary encoder is one of the reliable and cheap method in terms of room condition. Most of algorithms for a self localization are based on a fully floating point operation and trigonometry functions. Because of these reasons high-performance controllers should be applied. We propose self-localization calculating algorithm for low-performance controller without using operations mentioned above. This algorithm uses only integer add and subtraction operations. Algorithm fails more in long distance. Experimental results for the simulation presented.
Key words: Differential wheeled mobile robot, self localization, localization, autonomous robot,

1. Background

Real time localization is a most important for autonomous robot control. Main purpose of paper is to calculate location of differential driving mobile robot with low-performance controller –especially 8 bit AVR. Former calculations use the location which we give. During this function robot shifts goes different way cause of track slippery and curve and errors can't be defined. Solution of this problem is that robot calculates its own location. Furthermore mobile robot fixes errors caused of track while robot is moving. Although we made an experiment on math-based GAZ algorithm with low-speed microcontroller errors occur when robot moves fast. We have been doing GAZ algorithm experiment on ARM processor. Also we implemented following algorithm.

2. Theory

Rotary encoder transmits data in the impulse from to controller. Length of the path is defined by amount of impulse transmitted. Encoder that we are using transmits 500 impulse when rotates once. Encoder wheel is 20 cm in diameter which means robot goes 0.04 cm in one impulse. Figure 1 shows encoder data transmission.

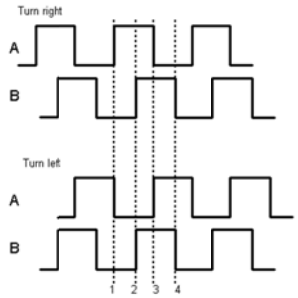


Figure 1. Data transmission of rotating encoder

Encoder transmits two series of impulse as shown in Figure 1. When first impulses interrupt occurs decoder defines wheel rotation direction backward or forward if next impulse is high or low respectively.

3. Modeling

We modeled the differential driving mobile robot's single step as shown in the following figure 2. There is no need to calculate precisely because an accuracy of the rotary encoder we are using is not good enough. Mobile robot's single step created tiny-length arc so the arc can be modeled as straight line. This way our estimate will be easy. A Beta angle, as shown in Figure 2, is the angle which is created then the robot's right wheel crosses one impulse or l = 0.04 cm path. The L is a distance between two rotary encoder wheels.

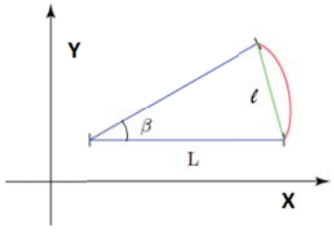


Figure 2. Model of rotating encoder

If mobile robot's two wheels go straight with different velocity, robot will deviate due to difference of velocity. Hence we can calculate robot's angle by using the difference of two rotating encoders. Also we are able to calculate robot's coordinate on every step of robot.

4. Algorithm

The main form of algorithm shown by (1) and (2).

$$X = X_0 + X(\beta) \quad (1)$$
$$Y = Y_0 + Y(\beta) \quad (2)$$

X −Present x axis value
Y −Present y axis value
X_0 −Previews x axis value
Y_0 −Previews y axis value
$X(\beta)$ −X axis projection value
$Y(\beta)$ − Y axis projection value

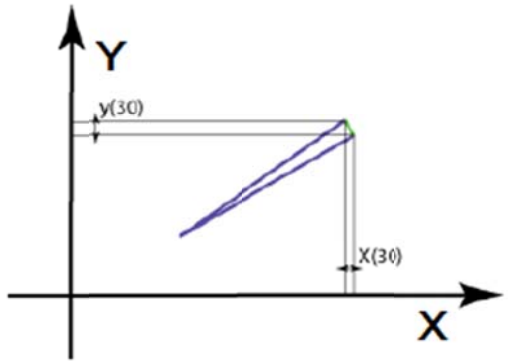


Figure 3. Robot movement of one step

It means add x and y axis's projections of one step of rotary encoder on present X, Y of location for determining next X, Y.

$$X(\beta) = l * sin(\beta) \quad (3)$$
$$Y(\beta) = l * cos(\beta) \quad (4)$$

β − Mobile robot's angle
l − Rotary encoder one step length

It only calculates on one rotary encoder wheel of robot. Because we can determine any other point's location of robot by robot angle, dimensions between the rotary encoder and point. But this calculation needs high performance, because of trigonometry functions and floating point mutilations.

5. Implementation

A losing resources (time) of this algorithm is that calculating (form (3) and (4)) floating point mutilations and trigonometry functions on every rotary encoder interrupt. But result of that operations are not a dynamic. You can see from figure 4 –how the coordinates X, Y will be changed depending on the mobile robot's angle on coordinate plane.

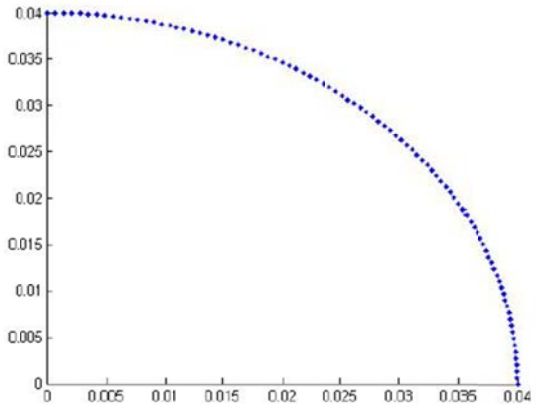


Figure 4. Single step correspoding values

So we can use static array consist of single step corresponding values on $l = 0.04$ cm. It uses memory resource, not timing resource.

$$X(n(\gamma)) = l * sin(\beta) \quad (5)$$
$$Y(n(\gamma)) = l * cos(\beta) \quad (6)$$

β − Mobile robot's angle
l − Rotary encoder one step length
n − Table index

The main algorithm represented by following.

$0 \leq \beta < 90$,
$$X = X_0 + x(\beta), Y = Y_0 + y(\beta) \quad (7)$$
$90 \leq \beta < 180$,
$$X = X_0 - x(\beta), Y = Y_0 + y(\beta) \quad (8)$$
$180 \leq \beta < 270$,
$$X = X_0 - x(\beta), Y = Y_0 - y(\beta) \quad (9)$$
$270 \leq \beta < 360$,
$$X = X_0 + x(\beta), Y = Y_0 - y(\beta) \quad (10)$$

β −Robot angle
X −Present x axis value
Y −Present y axis value
X_0 −Previews x axis value

Y_0 −Previews y axis value
$X(\beta)$ −X axis projection value of table
$Y(\beta)$ − Y axis projection value of table

Robot walks one step, when angle of Robot is 30, additional value of X, Y and angle must be known. To find angle of Robot is difference between encoder of two wheels. The angle of one impulse is dividing 360 to number of impulse to circle around.

$$\alpha = \frac{360}{n} \qquad (11)$$

n –number of impulse to circle around

An average L of robots that we use is approximately 50 cm. Hence beta angle change slowly comparative one step of rotary encoder.

6. Experiment

We made mobile robot model in the Matlab for experiment to test. We compared with GAZ algorithm which is close to test on the real time processor. The green line is result of our experiment. But the red and blue line is result of GAZ algorithm. We tested it on few other paths and made some conclusion. We draw eight number and result is in the figure 5 and 6.

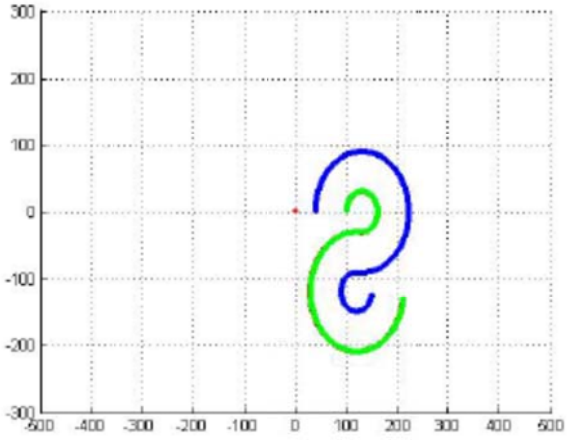


Figure 5. Number eight path

Error of 7100 impulse, on the other hand 284 cm path is 1 cm for X and 1cm for Y. You can see the result of walking on the straight path which is in the figure 7, 8. For straight path, our algorithm's error is so little. This is because of going for shorten. Error of 100000 impulse, on the other hand 4000 cm path is 1 cm for X and nothing for Y. For curved path, you can see result in the figure 9 and 10.

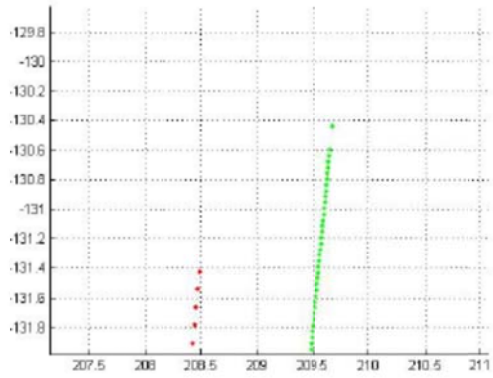


Figure 6. Zooming of number eight path

Error of 7000 impulse, on the other hand 280 cm path is 2.5 cm for X and 1 cm for Y.

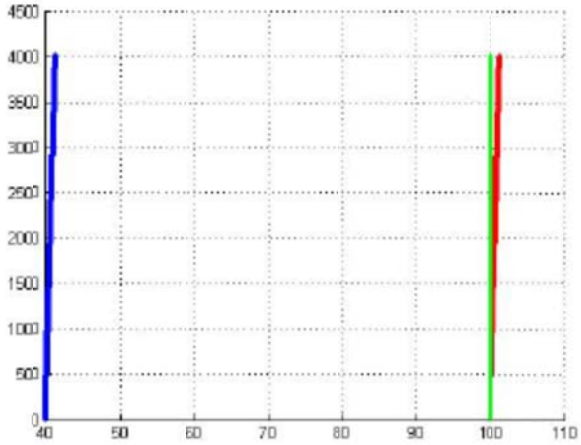


Figure 7. Straight path

We will save our string as an non character 32 bit, which is we can describe 4'294'967'295. We can make our calculation that can be 42 meter. Function that calculates the angle is same as mathematical function algorithm, so we didn't compare 2 of them. That is 1248 cycle; on the other hand it is 160 us in 14MHz. Maximum velocity to circle is 6.25 RPS, then our calculation will be true. Duration of our algorithm's performance is 102 cycles which is 13.3 us. For maximum value, we can calculate 30 m/s for steps of mobile robot.

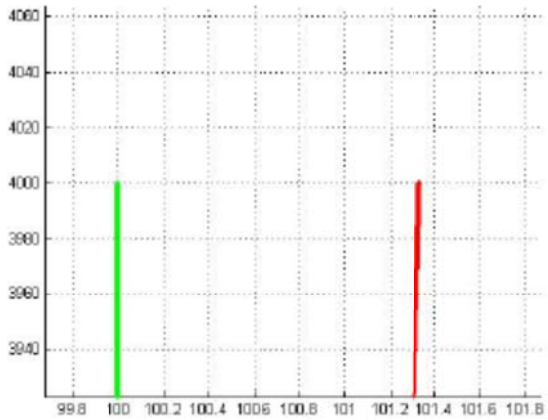


Figure 8. Zooming of straight path

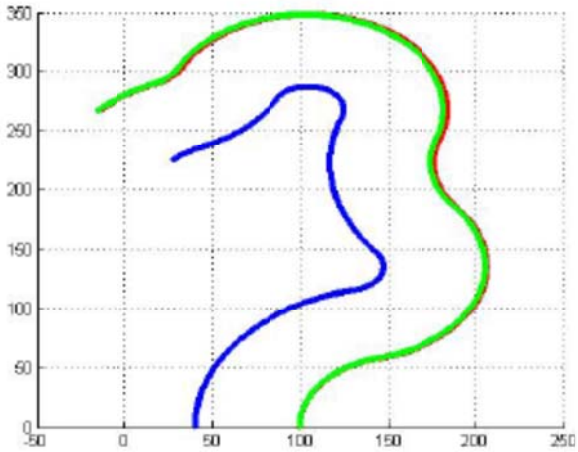


Figure 9. Curve path

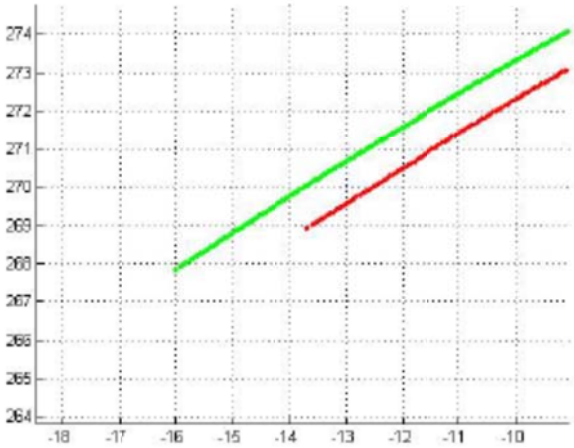


Figure 10. Zooming of curve path

Table 1. Error comparison

№	Type of path	Length of path(cm)	Error(cm)
1	Number 8 path	284	X=1, Y=1
2	Straight path	4000	X=1, Y=0
3	Curve path	280	X=2.5, Y=1

7. Conclusion

This method is easy to use and which requires low performance controller to find location of mobile robot. This algorithm requires angle of the robot must be true and will use floating point operation. But it is slow comparing to the degree movement robot's value and won't be any problem to calculate floating point operation. We tested error on the Matlab program. Time for obtain robot's location is 10 times faster than method of floating point operation. So we can use low performance controller.

8. Reference

[1] Ganbat G, Nanzadragchaa D, Bayarpurev M. "Differential wheeled robot self localization algorithm based on timing information", *MMT 2013*, 2013, pp.103 – 108.

[2] Ojeda L. and Borenstein J. "Methods for the reduction of odometer errors in over-constrained mobile robots", *Autonomous Robots*, Vol.16, No.3, 2004, pp.273-286.

[3] Edouard Ivanjko, Ivan Petrovic and Nedjeljko Peric. "An approach to odometer calibration of differential drive mobile robots", *In International Conference on Electrical Drives and Power Electronics*, 2003, pp.519–523.

Session : Special Keynotes and Banquet

- Triple Helix Enhancing Innovation

  Sermkiat Jomjunyong

- Policies for Promoting Basic Convergence Research by the National Research Foundation(NRF) of Korea

  Cha Eun Jong

Policies for Promoting Basic Convergence Research by the National Research Foundation (NRF) of Korea

Cha Eun Jong
National Research Foundation (NRF) of Korea
ejc6331@nrf.re.kr

Abstract

The NRF was established in 2009 to balance among different academic areas and to create an efficient support system for basic research. More than 500 people are serving for NRF and this year (2014) budget reaches 3,572 million US dollars. To break down any disciplinary barriers, a mission was announced in 2013 to nurture a convergence research friendly ecosystem. Convergence research areas were re-categorized into 5 and supported 103 US million dollars. A 5 step convergence research review process was also introduced to select best proposals. The NRF commits herself to promotion of basic convergence research activities nationwide.

Session : Interactive Session 2

- Comparison of Prognosis Factors between ST-Segment Elevation Myocardial Infarction and non-ST-Segment Elevation Myocardial Infarction of Patients with Atrial Fibrillation

  Ho Sun Shon, Jang-Whan Bae, Byung Jun Cho, Young Sung Lee, Young Gyu Kim

- Short-Term Electricity Price Forecasting using Cascade Neural Network

  Cheng Hao Jin, Hyun Woo Park, Ling Wang, Kyung Hee Lee

- Ensemble Method based MicroRNA Selection for Disease Diagnosis

  Minghao Piao, Yongjun Piao, Feifei Li, Keun Ho Ryu

- The Construction of Integration Databset for Correlation Analysis of Heart disease and Meteorological Information

  Hyeongsoo Kim, Kwang Sun Ryu, Jae Won Lee, Kwan Hee Yoo

- The Generation of Fusion factor for Acute Myocardial Infarction based on Causal Association Rule Mining

  Kwang Sun Ryu, Seung Hyeon Yang, Hyun Woo Park, Soo Ho Park, Ibrahim M. Ishag, Jang Whan Bae

- Biomedical Event Extraction with Random Forests

  Tsendsuren Munkhdalai, Meijing Li, Khuyagbaatar Batsuren, Wan-Sup Cho

- CUDA-based Multiple Linear Regression for Analysis of Large Health Data

  Soo Ho Park, Ho Sun Shon, Eun Jong Cha

Session : Interactive Session 2

- A Progressive Architecture for Source Code Clone Detection and Extraction by Using Data Ming Methods and MapReduce Paradigm
  Dingkun Li, Minghao Piao, Jong Yun Lee
- Correlation analysis of RNA-Seq and qRT-PCR for detection of differentially expressed genes
  Yongjun Piao, Nak Hyeon Choi, Meijing Li, Keun Ho Ryu
- Gait identification using Wearable Sensors with Spatio-Temporal Features
  Hyun Woo Park, Kyeong Seok Lee , Soo Ho Park, Cheng Hao Jin,
- Development of System to Measure Length of Moving Object on CCTV Image
  Kyu Ik Kim, Sunny Song, Myung-Sic Kim, Jin Suk Kim
- Accelerating the morphology operations using a CUDA on Graphics Processing Units
  Amartuvshin Renchin-Ochir, Gerelttulga Galaa, Bolormaa Dalanbayar
- Programmable logic chip based Sinewave generator using Cordic algorithm
  Battogtokh.J, Zorig.B, Bolormaa.D
- Technical analysis on 3D reconstruction methods
  Erdenebat Budsuren, Baatarbileg Altangerel, Oyun-erdene Namsrai

Comparison of Prognosis Factors between ST-Segment Elevation Myocardial Infarction and non-ST-Segment Elevation Myocardial Infarction of Patients with Atrial Fibrillation

Ho Sun Shon[1], Jang-Whan Bae[2], Byung Jun Cho[1], Young Sung Lee[1,] Young Gyu Kim[1]
[1]*Graduate School of Health Science Business Convergence, Chungbuk National University*
{shon0621, cho135135, lee.medric, brsurg}@gmail.com
[2]*Department of Medicine, School of Medicine, Chungbuk National University*
jangwhanbae69@gmail.com

Abstract

Background: A trial fibrillation means irregular heartbeat as the most common arrhythmia. Despite the importance of atrial fibrillation in acute myocardial infarction, it has currently not defined for antiarrhythmic drug, kinds of inserted stent, medicine for regulating appropriate heart rate, factor influencing to long-term prognosis. In this paper, we compare and analyze STEMI (ST-segment elevation myocardial infarction) patients and NSTEMI (non-ST-segment elevation myocardial infarction) patients among the patients with acute myocardial infarction and atrial fibrillation in Korea, find out the factor with significant difference, and intend to help classification of the patients.

Methods: Among 14,886 patients registered with acute myocardial infarction in Korea from November of 2005 to March of 2008, the target is 574 patients (male : female = 382 : 189, 77±12 years) with atrial fibrillation. Among them STEMI patients are 300, and NSTEMI are 268.

Results: The incidence rate of atrial fibrillation in acute myocardial infarction is 3.8% (574/14,886), the incidence rate of major cardiac event is 3.1% (18/574). In comparison between STEMI and NSTEMI, there seems no odds in diastolic blood pressure (p=0.366), total cholesterol (p=0.202), and so on. But there is significant difference in biomarker, systolic blood pressure (p=0.001), heart rate (p=0.001), Glucose (p=0.028), NT-proBNP (p=0.008). There is significant difference in every step of Killip class in multivariate analysis of NSTEMI.

Conclusions: In case of having atrial fibrillation in acute myocardial infarction in Korea. Glucose, NT-proBNP, and Killip class are predictive factor for classifying STEMI and NSTEMI. Especially, it is possible to classify in every step of Killip class. Also it is possible to help diagnosis and treatment for the patients using this theory.

Acknowledgement

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

Short-Term Electricity Price Forecasting using Cascade Neural Network

Cheng Hao Jin[1], Hyun Woo Park[1], Ling Wang[2], Kyung Hee Lee[3]
[1]*Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*
{kimsungho, hwpark}@dblab.chungbuk.ac.kr
[2]*Department of Computer Science and Technology, Northeast Dianli University, China*
smile2867ling@gmail.com
[3]*Department of Business Data Convergence, Chungbuk National University, Korea*
lee.kyunghee@gmail.com

Abstract

Electricity price forecasting is an important issue in competitive electricity markets. A lot of techniques have been used for forecasting the electricity price. However, to the best of our knowledge, there is little work on forecasting electricity price with Cascade 2 training algorithm. In this paper, we present its forecasting results performed on real-world Spanish electricity price datasets. The experimental results that the cascade neural network with cascade 2 training algorithm could achieve satisfactory results in electricity price forecasting.

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2008-0062611) and supported by the MSIP(The Ministry of Science,ICT and Future Planning), Korea, under the "SW master's course of a hiring contract" support program (NIPA-2014-HB301-14-1011) supervised by the NIPA(National IT Industry Promotion Agency).

Ensemble Method based MicroRNA Selection for Disease Diagnosis

Minghao Piao, Yongjun Piao, Feifei Li, Keun Ho Ryu*
Database/Bioinformatics Laboratory, Department of Computer Science,
Chungbuk National University, South Korea
{bluemhp, pyz, feifeili, khryu}@dblab.chungbuk.ac.kr

Abstract

For the selection of most significant microRNAs and its use in human cancer classification, traditional feature selection methods are widely used like filter approach, embedded approach and wrapper approach. The weakness of these methods is that it would decrease the stability of biomarkers. Recently, ensemble feature selection methods are very popular in bioinformatics to improve the stability of biomarkers. However, there are still no studies about application of ensemble feature selection in microRNAs based human cancer classification. In our study, we describe a Cascading-and-Sharing and 10-fold cross validation based ensemble feature selection of microRNAs and human cancer classification. The results show that our approach can select most significant microRNAs with high quality of classification.

1. Ensemble Feature Selection

When looking for biomarkers from DNA, RNA and microRNAs expression data, only a small subset of biomarkers are selected which are related to specific diseases. One of the most common approaches is through ordering or ranking the genes by its importance. Ordering genes by its importance is very similar to feature selection which is a preprocessing step of data mining. The feature selection methods can return a set of features that are most important to the problem at hand. Feature selection methods can be applied to several issues in biological and genetics: distinguishing between healthy and diseased tissue [1, 2, 3, 21]; identification and classification of different types of cancer [4, 5]; prediction of drug treatment [6, 7], etc.

*Corresponding author

In bioinformatics, data sets usually contain few samples (often less than a hundred) and thousands of different genes (curse of dimensionality). This will decrease the stability of feature rankers and lead to generating different results after slightly changing the data set [8].

In order to improve the stability of feature selection techniques, researchers have proposed new frameworks such as ensemble feature selection methods [9-15]. The idea for ensemble feature selection is derived from ensemble learning methods wherein different classifiers are applied to a dataset and their results are aggregated. Ensemble feature selection techniques apply feature selection algorithms multiple times and combine the results into the decision making. Because combining multiple results, the features which are frequently chosen as the best performers will be marked as top-ranked features, while features with poor performance will be low-ranked features; thus, the final top-ranked features will be more stable. There are three main types of ensemble feature selection techniques [16]. (1) Data diversity consists of applying a single feature selection method to a number of differently sampled versions of the same dataset and then an aggregation technique is used to aggregate the results. (2) Functional diversity is performed by applying a set of different feature selection techniques on the same dataset. (3) Hybrid ensembles use both of these, applying different feature selection techniques to different sampled versions.

2. Decision Tree Ensemble based Feature Selection

In data mining, ensemble methods are used for improving the classifier's accuracy. Ensemble methods are used to construct a set of base classifiers from training data set and perform the classification work by voting on the predictions made by each classifier. Since the idea of ensemble feature selection is derived

from ensemble learning methods, it is possible to apply ensemble learning methods in feature selection if the method can decide which features to construct the set of classifiers.

The ensemble of classifiers can be constructed in many ways [17] and most widely used is by manipulating the training set like Bagging and boosting. Three interesting observations are described in [18] based on the study of many ensemble methods: (1) Many ensembles constructed by the Boosting method were singletons. Due to this constraint, deriving classification rules have a limitation: decision trees are not encouraged to derive many significant rules and they are mutually exclusive and covering the entire of training samples exactly only once. (2) Many top-ranked features possess similar discriminating merits with little difference for classification. This indicates that it is worthwhile to employ different top-ranked features as the root nodes for building multiple decision trees. (3) Fragmentation problem is another problem does those ensemble methods have: as less and less training data are used to search for root nodes of sub-trees.

Base on those observations, if we want to apply ensemble learning method to feature selection in bioinformatics, we need a method that can break the singleton coverage constraint and solve the fragmentation problem. Cascading-and-Sharing is proposed to solve these problems [18, 19].

In this study, Cascading-and-Sharing is used as feature selection techniques, and K-fold cross validation is used to produce differently sampled versions of the same dataset. Cascading-and Sharing method is applied on these datasets to compose the Data diversity (ensemble feature selection) of microRNAs and human cancer classification.

3. Experimental Results

3.1 microRNA Dataset

The microRNA expression dataset was first published by [20]. They used a bead-based method to present a systemic expression analysis of 217 mammalian microRNAs from 186 samples including multiple human cancers. The used data set in this paper is described in Table 1.

Table 1. The number of samples for each cancer type

Cancer Name	No. of Tumor Samples
Colon	10
Pancreas	9
Uterus	10
Mesothelioma	8
Breast	6
B Cell ALL	26
T Cell ALL	18
Follicular Cleaved Lymphoma	8
Large B Cell Lymphoma	8
SUM	103

3.2 Feature Selection and Classification

The CS4 algorithm with different number of top-ranked microRNAs is running on the datasets produced by 10-cross validation. The top-ranked microRNAs with higher performance is chosen and marked as most significant microRNAs. Figure 1 shows the classification accuracy of our approach on the 10 ~ 217 microRNAs with interval of 10. When the given top-ranked microRNAs are 10 ~ 60, the accuracy is increasing with bigger number of microRNAs. And, the classifier shows the higher accuracy when the given number of microRNAs is 60 and 80. Therefore, it is better to choose top-ranked microRNAs as most significant microRNAs in the interval of 50 ~ 80.

Figure 2 shows the classification accuracy of the method on the 50 ~ 80 microRNAs. We can see that the classifier shows higher accuracy when the given microRNAs are 55 ~ 64 and 73 ~ 80. When the given number of top-ranked microRNAs is 65 ~ 72, the accuracy is decreased. It means there are several microRNAs that are not suitable to build decision tree committees when the number is bigger than 64 even it shows higher accuracy on 73 ~ 80 microRNAs. Also, the cost of decision tree induction will become expensive when considering too many features during the process. From Table 2, we can see that the performances on 55 ~ 64 and 73 ~ 80 microRNAs are same. Therefore, we are choosing 55 top-ranked microRNAs as most significant microRNAs in human cancer classification.

4. Conclusion

In data mining, traditional feature selection methods can be divided into filter approach, embedded approach and wrapper approach. In bioinformatics, the weakness of these methods is that it would decrease the stability of biomarkers. Recently, ensemble feature selection methods are very popular in bioinformatics to improve the stability of biomarkers. The ensemble feature selection methods can be divided into three types: Data diversity, Functional diversity and Hybrid ensemble. However, there are still no studies about

application of ensemble feature selection in microRNAs based human cancer classification. In our study, we described a Cascading-and-Sharing and 10-fold cross validation based ensemble feature selection of microRNAs and human cancer classification. The experimental results show that our approach is useful to define most significant microRNAs by evaluating the classification performance of top-ranked features. Also, we have found several most common microRNAs which are most often used in decision tree induction with different number of top-ranked.

Our future work will be focusing on the application of Functional diversity and Hybrid ensemble method on microRNAs and trying to design new ensemble feature selection method.

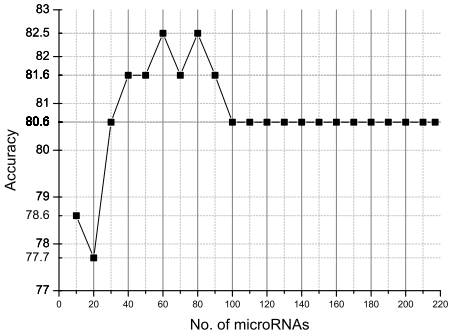


Figure 1. Classification accuracy on different number of top-ranked microRNAs

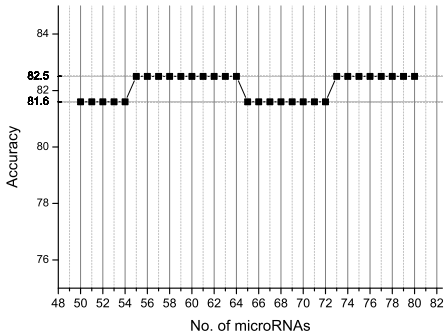


Figure 2. Classification accuracy on 50~80 top-ranked microRNAs

Table 2. Detailed classification performance

No. microRNAs	Classes (No. of instances)	Precision	Recall	F-measure
55 ~ 64	Colon (10)	0.778	0.7	0.737
	Pancreas (9)	0.778	0.778	0.778
	Uterus (10)	0.7	0.7	0.7
	Mesothelioma (8)	0.833	0.625	0.714
	Breast (6)	0.75	1	0.857
	B Cell ALL (26)	0.926	0.962	0.943
	T Cell ALL (18)	0.947	1	0.973
	Follicular Cleaved Lymphoma (8)	0.714	0.625	0.667
	Large B Cell Lymphoma (8)	0.625	0.625	0.625
73 ~ 80	Colon (10)	0.778	0.7	0.737
	Pancreas (9)	0.778	0.778	0.778
	Uterus (10)	0.7	0.7	0.7
	Mesothelioma (8)	0.833	0.625	0.714
	Breast (6)	0.75	1	0.857
	B Cell ALL (26)	0.926	0.962	0.943
	T Cell ALL (18)	0.947	1	0.973
	Follicular Cleaved Lymphoma (8)	0.714	0.625	0.667
	Large B Cell Lymphoma (8)	0.625	0.625	0.625

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 (NIPA-2014-H0301-14-1022) supervised by the NIPA (National IT Industry Promotion Agency), and Basic Science Research

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

5. Reference

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proceedings of the National Academy of Sciences, vol. 96, no. 12, 1999, pp. 6745-6750.

[2] S. Dudoit, J. Fridlyand, and T. P. Speed, Comparison of discrimination methods for the classifi-cation of tumors using gene expression data, Journal of the American Statistical Association, vol. 97, no. 457, 2002, pp. 77-87.

[3] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning, vol. 46, 2002, pp. 389-422.

[4] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini, Tissue classification with gene expression profiles, Journal of Computational Biology, vol. 7, no. 3-4, 2000, pp. 559-583.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, Molecular classification of cancer:Class discovery and class prediction by gene expression monitoring, Science, vol. 286, no. 5439, 1999, pp. 531-537.

[6] D. Dittman, T. Khoshgoftaar, R. Wald, and A. Napolitano, "Random forest: A reliable tool for patient response prediction," Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM) Workshops. BIBM, 2011, pp. 289-296.

[7] G. Mulligan, C. Mitsiades, B. Bryant, F. Zhan, W. J. Chng, S. Roels, E. Koenig, A. Fergus, Y. Huang, P. Richardson, W. L. Trepicchio, A. Broyl, P. Sonneveld, J. Shaughnessy, John D., P. Leif Bergsagel, D. Schenkein, D.-L. Esseltine, A. Boral, and K. C. Anderson, Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib, Blood, 2007, pp. 3177-3188.

[8] A. Kalousis, J. Prados, and M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowledge and Information Systems, vol. 12, no. 1, 2006, pp. 95-116.

[9] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics, vol. 26, no. 3, 2010, pp. 392-398.

[10] A. C. Haury, P. Gestraud, and J. P. Vert, The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures, PLoS ONE, vol. 6, no. 12, 2011, pp. e28210.

[11] H. Liu, L. Liu, and H. Zhang, Ensemble gene selection by grouping for microarray data classification, Journal of Biomedical Informatics, vol. 43, no. 1, 2010, pp. 81-87.

[12] Y. Saeys, T. Abeel, and Y. Peer, "Robust feature selection using ensemble feature selection techniques," Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases - Part II. Berlin, Heidelberg: Springer-Verlag, 2008, pp.313-325.

[13] P. Yang, J. Ho, Y. Yang, and B. Zhou, Gene-gene interaction filtering with ensemble of filters, BMC Bioinformatics, vol. 12, no. Suppl 1, 2011, pp. S10.

[14] P. Yang, Y. Hwa Yang, B. B Zhou, and A. Y Zomaya, A review of ensemble methods in bioinformatics, Current Bioinformatics, vol. 5, no. 4, 2010, pp. 296–308.

[15] L. Yu, Y. Han, and M. E. Berens, Stable gene selection from microarray data via sample weighting, IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 9, no. 1, 2012, pp. 262–272.

[16] Awada, Wael, et al. "A review of the stability of feature selection techniques for bioinformatics data", 2012 IEEE 13th International Conference on Information Reuse and Integration (IRI), 2012, pp. 356-363.

[17] P. N. Tan, M. Steinbach, V. Kumar, Ensemble methods. Introduction to data mining, Addision Wesley, 2006, pp. 278-280.

[18] J. Y. Li, H. A. Liu, See-Kiong Ng, Limsoon Wong, Discovery of significant rules for classifying cancer diagnosis data, Bioinformatics, vol. 19, 2003, pp. 93-102.

[19] J. Li, H. Liu, "Ensembles of cascading trees", Proceedings of Third IEEE international conference on data mining, 2003, 585-588.

[20] E. Fridman, Z. Dotan, I. Barshack, M.B. David, A. Dov, S. Tabak, O. Zion, S. Benjamin, H. Ben-jamin, H. Kuker, Accurate molecular classification of renal tumors using microRNA expression, The Journal of molecular diagnostics, 2010, pp. 687-696.

[21] Y. J. Piao, H. W. Park, C. H. Jin and K, H. Ryu, "Ensemble Method for Classification of High Dimensional Data", Proceedings of the International Conference on Big Data and Smart Computing, 2014, pp. 245-249.

The Construction of Integration Dataset for Correlation Analysis of Heart disease and Meteorological Information

Hyeongsoo Kim[1], Kwang Sun Ryu[1], Jae Won Lee[2], Kwan Hee Yoo[2]
[1]*Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*
{hskim,ksryu}@dblab.chungbuk.ac.kr
[2]*Deptment of Business data and convergence, Chungbuk National University, South Korea*
inodie86@gmail.com, khyoo@cbnu.ac.kr

Abstract

Heart disease is the leading cause of death in the world over the past 10 years and its attack rate has been on the rise. A lot of studies for prevention of heart disease were progressed, and risk factors that affect in heart disease were published through study results. However, a study for various environmental factors that might cause heart disease is unprepared, because previous researches were studied based on patient's clinical information. Therefore, we aim to develop a prediction system to prevent outbreak of heart disease through correlation analysis with environmental factors and heart disease. As first step in our research, we describe the construction of an integrated database that can be basis of association analysis of heart disease and meteorological information using KAMIR and meteorological data. To integrate two sources of data, we classified 79 locations of meteorological sites and assigned a meteorological site to each patient's location information. Thereafter, we matched meteorological information with patient's arrival date to the hospital. We are expecting that correlation analysis between heart disease and meteorological information might reveal causal relationship.

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program(NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2008-0062611) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2013R1A2A2A01068923).

The Generation of Fusion factor for Acute Myocardial Infarction based on Causal Association Rule Mining

Kwang Sun Ryu[1], Seung Hyeon Yang[1], Hyun Woo Park[1], Soo Ho Park[1], Ibrahim M. Ishag[1], Jang Whan Bae[2]

[1]*Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*

{ksryu, disroway, hwpark, soohopark, Ibrahim}@dblab.chungbuk.ac.kr

[2]*Department of Medicine, School of Medicine, Chungbuk National University, South Korea*

jangwhanbae69@gmail.com

Abstract

Cardiovascular diseases are still increasing in modern societies due to the adaptation of western life, smoke and obesity. Especially, the acute myocardial infarction (AMI) which shows higher mortality compared with other cardiovascular diseases. Following this trend, research of AMI is being carried out actively to find the factors affecting its outbreak. However, existing factors lack predictive power raising a demand for more representative factors. Therefore, we propose a novel fusion factor (FF) in order to diagnosis AMI using causal association rule mining. We expect that the FF should support the prediction of AMI effectively.

Acknowledgement

This work 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) and NRF grant funded by the Korea government (MSIP) (No. 2008-0062611).

Biomedical Event Extraction with Random Forests

Tsendsuren Munkhdalai[1], Meijing Li[1], Khuyagbaatar Batsuren[1], Wan-Sup Cho[2]
[1]*Database/Bioinformatics Laboratory, Department of Computer Science,*
Chungbuk National University, South Korea
{tsendeemts, mjlee, huygaa }@dblab.chungbuk.ac.kr
[2]*Dept. of MIS/Business Data Convergence,*
Chungbuk National University, Cheongju, Korea
wscho@chungbuk.ac.kr

Abstract

As biomedical literature on servers grows exponentially in the form of semi-structured documents, biomedical text mining has been intensively investigated to find information in a more accurate and efficient manner. Extraction of biomedical event has received recent attention, as it allows uncovering the complex structured knowledge from biomedical event descriptions written in text data.

In this paper, we introduced a pipeline approach with more complex classification model for biomedical event extraction. The event extraction task is decomposed into five different stages, preprocessing, named entity recognition, trigger detection, edge detection and event construction. The each stage other than the preprocessing one becomes a machine learning classification problem. We use random forests for the main classification problem by applying feature weights of SVM classifier as feature selection criterion while most of the previous approaches are based only on a simple linear classification model, namely SVM with Linear kernel because of the high dimensionality of the dataset.

During the experimental analysis, we notice that higher performance level is achievable with more complex model than the linear models, even though we were not able to train the random forests with more than 10K features on a single machine. Future research might be in a direction where the multiple machines are used as cluster for the building such a complex model with more features to incorporate necessary knowledge.

Acknowledgment

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) and by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency).

CUDA-based Multiple Linear Regression for Analysis of Large Health Data

Soo Ho Park[1], Ho Sun Shon[2], Eun Jong Cha[3]
[1]*Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*
soohopark@dblab.chungbuk.ac.kr
[2] *Graduate School of Health Science Business Convergence, Chungbuk National University, Korea*
shon0621@gmail.com
[3]*Department of Biomedical Engineering, Chungbuk National University, Korea*
ejcha@chungbuk.ac.kr

Abstract

In statistics, regression analysis is estimating the relationships among variables. Many applications of regression analysis involve more than two independent variables. A regression model that contains more than two independent variables is called Multiple Linear Regression(MLR) model. MLR is one of the most famous predictive method in many various fields including the medicine area. However, as the size of medical data has increased in recent years, the traditional algorithm would be unsuitable or spend a lot of time for handling this large dataset. Therefore, it is need to find solution to address these problems. One solution is the parallel computing technique on Graphics Processing Unit(GPU). GPU is originally designed to compute computer graphics only. Nowadays, GPU is widely used not only graphics rendering but general computation in many application domains. Today's GPU has more than hundreds of cores, it can be used to solve large problem rapidly using thousands of threads. Compute Unified Device Architecture(CUDA) is a parallel computing platform and programming model which is include compiler, debugger and graphics card driver by NVIDIA. It provides some CUDA-accelerated library and standard C-like interface to developer to help easily use parallel programming on CUDA GPUs.

In our study, we implements a CUDA based MLR algorithm to predict prognosis of cancer patients. To estimate the regression coefficients in MLR, matrix based least square estimation is used. CUDA basic Linear Algebra Subroutines(cuBLAS) and GPU accelerated linear algebra library are used to compute the matrix operation, because of there are many matrix operations in least square estimation such as matrix multiplication, matrix transpose and inverse matrix. The results shows that CUDA based parallel MLR is suitable in our study for prediction of prognosis of cancer patients from large cancer. In the future, we consider many various statistics method and data mining techniques and optimize that to analyze huge amount of data.

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) and the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency).

A Progressive Architecture for Source Code Clone Detection and Extraction by Using Data Ming Methods and MapReduce Paradigm

Dingkun Li[1], Minghao Piao[1], Jong Yun Lee[2]
[1]*Database/Bioinformatics Laboratory, Department of computer science,*
Chungbuk National University, South Korea
{jerryli, bluemhp}@dblab.chungbuk.ac.kr
[2]*Department of Digital Informatics and Convergence,*
Chungbuk National University, South Korea
jongyun@chungbuk.ac.kr

Abstract

With the development of the large software systems, the software maintain is becoming more and more difficult. One reason is the discretionary copy-paste source code blocks (called code clones) which greatly decreases the quality of the whole systems, causes a lot of error prone and time consuming effect during software maintenance phase, and obeys the encapsulation rule of the software engineering. In this paper we briefly introduce a progressive architecture for token-based source code clone detection and extraction by using data mining (DM) methods and MapReduce paradigm. Preprocessing, the first stage of research has already been done to support our further research. And an exploratory research by using MapReduce paradigm for parallel source code detection will be demonstrated and the DM methods for clones classification and clustering will be briefly introduced in this paper.

1. Introduction

The maintenance of large scale software projects are becoming more and more difficult, one reason is because these projects consist large amount of the discretionary and unmarked copy-paste source code blocks which can be called as source code clone among hundreds of thousands lines of code (LOC). When a fraction of source code block need to be modified, all the other clone code need to be located and modified as well. It is a time consuming and error prone process. Though these clones save a lot of time for programmers, it greatly decreases the quality of the whole systems, and obeys the encapsulation rule of the software engineering. It is important and necessary to locate these clones, classify them and encapsulate them in the super classes or global functions.

There are a lot of challenges to detect and extract source code clones from lager scale software systems.

The first problem is that these systems often consist hundreds of thousands LOC that clones are buried among the sea of code, the second challenge for source code clone detection and extraction is the preprocessing of the source code as it is unstructured data which contains a lot of irrelevant data such as comments, repeating statements, embedded code etc. The third challenge is clone detection because it is hard to figure out where is the start and where is the end of one clone block and how many clones and how many groups (or clusters) of clones the systems actually have. The third challenge is clone extraction, after all clones having been detected, it is hard to group these clones and figure out how similar the clones within the same group are. Also there are many other challenges such as what is a good way to calculate the similarity between two clones, what kind of the data structure can be used to record the clone groups and so on.

Data mining and MapReduce techniques provide a promising way to detect and extract the source code clones from large scale software systems. In this paper we provide a progressive architecture for source code clone detection and extraction by using these techniques and give an exploratory research. The first stage of research has already been done by our previous work [1]. This paper just gives guidance for further research.

2. Related Work

This section concludes some of the related work and important methods or concepts.

2.1 Code Clone

Code fragments: is a contiguous piece of source code [2], usually concludes several lines (more than one line in our work) of code. Sometimes it is also called statements. In [2] it said that clone consisting out of more than 5 statements are considered interesting.

Code clone: is a copy or several copies of a code fragment. Generally, code clone can be categorized into different types [3] which are listed in Table 1.

Table 1. Code clone type

Type	Comments
1	Identical code fragments except for variations in whitespace, layout and comments.
2	Syntactically identical fragments except for variations in identifiers, literals, types, whitespace, layout and comments.
3	Copied fragments with further modifications such as changed, added or removed statements, in addition to variations in identifiers, literals, types, whitespace, layout and comments.
4	Two or more code fragments that perform the same computation but are implemented by different syntactic variants.

Code clone can be divided into different levels [4] which are listed in Table 2.

Table 2. Code clone level

Level	Comments	
1	Repeating groups of simple clones	
	A	In the same method
	B	Across different methods
2	Repeating groups of simple clones	
	A	In the same file
	B	Across different files
3	Method clone sets	
4	Repeating groups of method clones	
	A	In the same file
	B	Across different files
5	File clone sets	
6	Repeating groups of file clones	
	A	In the same directory
	B	Across different directories
7	Directory clone sets	

In the work [2], types 1~3 from belong to simple clones.

The ways to detect the code clones can be generally classified into 6 categorizes such as String-based, Token-based, Parse trees-based, Program dependency graphs-based, Metrics-based and Hybrid approach.

2.2 Preprocessing

In real world of data, the quality of data always cannot satisfy the requirements of data mining. It always suffers from the noise, incompletion, inconsistence, and too many dimensions. These problems come from all aspects of the lifetime of data management, like obtaining, transmission, and storing [7].

So we need preprocessing step to enhance the quality of data, and then to enhance the efficiency and validation of data mining so as to get mining results with good quality. There are several major tasks of preprocessing:

- **Data cleaning** routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
- **Data** reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.
- **Data transformation** consists of normalization, data discretization, and concept hierarchy generation.
- **Data integration** merges data from multiple sources into a coherent data store such as a data warehouse.

These techniques are not mutually exclusive; they may work together. For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

2.3 Related Data Mining Techniques

Three techniques can be used for our work.

Frequent patterns are patterns (e.g., itemsets, subsequences, or substructures) that appear frequently in a data set. It searches for recurring relationships in a given data set.

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute.

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects

within a cluster have high similarity, but are very dissimilar to objects in other clusters.

2.4 MapReduce Paradigm

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster [8,9].

The Map and Reduce functions of MapReduce are both defined with respect to data structured in (key, value) pairs. Map takes one pair of data with a type in one data domain, and returns a list of pairs in a different domain:

Map(k1,v1) → list(k2,v2)

The Map function is applied in parallel to every pair in the input dataset. This produces a list of pairs for each call. After that, the MapReduce framework collects all pairs with the same key from all lists and groups them together, creating one group for each key.

The Reduce function is then applied in parallel to each group, which in turn produces a collection of values in the same domain:

Reduce(k2, list (v2)) → list(v3)

Each Reduce call typically produces either one value v3 or an empty return, though one call is allowed to return more than one value. The returns of all calls are collected as the desired result list.

Thus the MapReduce framework transforms a list of (key, value) pairs into a list of values. This behavior is different from the typical functional programming map and reduces combination, which accepts a list of arbitrary values and returns one single value that combines all the values returned by map.

It is necessary but not sufficient to have implementations of the map and reduce abstractions in order to implement MapReduce. Distributed implementations of MapReduce require a means of connecting the processes performing the Map and Reduce phases. This may be a distributed file system. Other options are possible, such as direct streaming from mappers to reducers, or for the mapping processors to serve up their results to reducers that query them.

3. Main Framework and Methods

This section gives a brief introduction to the main framework and how the data mining methods can used for our purpose. Figure 1 shows the work flow of the main framework.

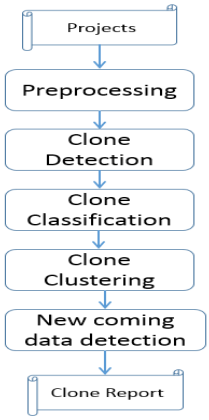


Figure 1. Progressive architecture of the system

3.1 Preprocessing

In data mining area, including source code mining, preprocessing plays an important role for further processing. Just like the old saying goes: well begun is half done. The output of the preprocessing---tokens are the most important data for clone detection. In our work [1], we provide a novel way to preprocess the source code files of the large scale software systems such as Linux, Wildfly, VTK etc. All the words, numbers, punctuation marks, parentheses and quotation marks are considered as tokens. The tool we developed called OPP (one pass preprocessor) integrates all the tokens into a single txt file for mining. The experiment performs well as is shown in table 3.

Table 3. Case study system result

System	Directory number	Accepted File number	CPU time(ms)
Wildfly1.02	4509	8787 (java)	21984
Linux core-3.6	2477	17448 (c)	297778
VTK	1116	4556 (c++)	49334

3.2 Clone Detection

In clone detection step, first we define some problem formulations: A project consists of blocks of source code P: {B1, B2, B3, ...} A block of source code consists of various source terms B: {T1, T2, T3, ...}. Given two projects P1 and P2, a similarity function f, and a threshold parameter t, the aim is to compute and find all block pairs P1.B and P2.B where $f(P1.B, P2.B) \geq T$. The principle idea of the approach is to hash partition the source code blocks across the network based on the computed keys and group together source code blocks which have the same key.

In our work, we use the output of the previous step as the input of this step. For each block B it is one line of the output file. For each token T it is one token of B.

Then we use map-reduce procedure three times: **for the first time**, the mapper produces a key-value pair of the form<T,1> for each term in B. The reducer computes the total occurrence for each source term and output<T, value> key-value pairs, where value is the total frequency for the term across all blocks. **For the second time**, the map function exchanges the input keys and values so that the input pairs of the reduce function are sorted based on their frequencies. The reducer just outputs the values without keys, the output is called as rarity term list. **For the third time**, the mapper retrieves the original blocks one by one, tokenizes it and reorders the tokens based on their frequencies, using the rarity term list as a reference. Next, it computes the prefix length and extracts the prefix terms (The basic idea behind prefix terms principle is that if two blocks of source code share rare terms, there is a chance that it might be similar). Considering each source term as a key, we produce <key, value> pair for each of its prefix terms. Hence we project a source code block multiple times (the number of its prefix terms). The reducer groups together all the values sharing same prefix tokens. In the end, a single reducer, for each pair of code block projections applies length and positional filters and checks the pair if it survives. If a pair passes the similarity threshold set by the user, the reducer outputs block pairs as clones with their similarity values.

In our work, we plan to setup cluster using Amazon web services with the following node speci- fication: 1.7 GB memory, 160 GB storage, 1 EC2 compute unit (1 virtual core), Hadoop 0.20. We speculate performance gain to improve with larger dataset (order of 100 and 1000 of projects) as HDFS performs better with large files.

3.3 Clone Classification

Then classifying methods of data mining will be explored to classify the clones and put them into different clone pools. The classifier can be established during this process, and it is a self-improving process as more and more training data will be used to improve the precision of the classifier.

3.4 Clone Clustering

After that, clustering method of data mining will be used to improve the quality of each pool or cluster.

4. Conclusion

In this paper we briefly introduce a main framework for token-based source code clone detection and extraction. The first step has already been implemented by our previous work. The next few steps will be done in our next stage of research. Data mining techniques such as frequent pattern mining, classification and clustering have been explored to do support our purpose. The final goal of our work is to help improving the quality and productivity of software products.

Also we hope our source code clone detection techniques can be used to automatically detect new coming source code and give some suggestions to the programmers or software architect to reuse or abstract clone code. Finally we hope our tool can help to improve the quality and productivity of software products.

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2008-0062611) and performed as a subproject of project "Building Scientific Big Data Sharing and Convergence System" and supported by the Korea Institute of Science and Technology Information (KISTI) in 2014.

Reference

[1] Dingkun Li, Minghao Piao, Ho Sun Shon, Incheon Paik, Keun Ho Ryu, "One Pass Preprocessing for Token-based Source Code Clone Detection", not published.

[2] Vera Wahler, Dietmar Seipel, Jürgen Wolff v. Gudenberg, and Gregor Fischer, "Clone Detection in Source Code by Frequent Itemset Techniques", *Fourth IEEE International Workshop on Source Code Analysis and Manipulation, Chicago, USA*, 2004, pp.128-135.

[3] Roy C K, Cordy J R, Koschke R, "Comparison and evaluation of code clone detection techniques and tools: A qualitative approach", *Science of Computer Programming*, 74(7), 2009, pp.470-495.

[4] Basit H A, Jarzabek S, "A data mining approach for detecting higher-level clones in software", *Software*

Engineering, IEEE Transactions on, 35(4), 2009, pp.497-514.

[5] Yan X, Han J, Afshar R, "CloSpan: Mining closed sequential patterns in large datasets", Proceedings of SIAM International Conference on Data Mining, 2003, pp.166-177.

[6] Zaki M J, "SPADE: An efficient algorithm for mining frequent sequences", *Machine learning*, 42(1-2), 2001, pp.31-60.

[7] J.W Han, M. Kamber, and J. Pei, "Data Mining: Concept and Techniques, 3rd ed", Morgan Kaufmann, 2011, pp.332-350.

[8] Google spotlights data center inner workings | Tech news blog - CNET News.com

[9] Dean J, Ghemawat S, "MapReduce: simplified data processing on large clusters", *Communications of the ACM*, 51(1), 2008, pp.107-113.

Correlation analysis of RNA-Seq and qRT-PCR for detection of differentially expressed genes

Yongjun Piao, Nak Hyeon Choi, Meijing Li, Keun Ho Ryu
Database/Bioinformatics Laboratory, Department of computer science,
Chungbuk National University, South Korea
{pyz, nak, mjlee, khryu}@dblab.chungbuk.ac.kr

Abstract

Recently, the emergence of next-generation sequencing (NGS) technology brings significant changes in many biological applications. Whole transcriptome shotgun sequencing, also known as RNA-Seq, is rapidly become the standard method for transcriptomics. In the past year, a lot of researches have been done utilizing RNA-Seq for quantifying expression levels, detection of alternative splicing, and discovery of novel transcripts. However, the primary objective of many biological studies is to identify differentially expressed genes in different conditions. For the past decade, microarrays have been widely used to simultaneously measure the expression levels of tens of thousands genes, but RNA-Seq is increasingly being used for quantification of expression levels of mRNAs as an alternative for microarray nowadays. It is because RNA-Seq has the advantage of i) low background noise, ii) dynamic range to quantify gene expression level, iii) ability to identify different isoforms, and iv) relative low cost compared with microarray.

Generally, RNA-Seq analysis pipeline for differential expression analysis consists of the following procedures. An RNA sample is converted to cDNA fragments or RNA fragments with adapters and sequenced on a high-throughput sequencing platforms, such as Illumina and Roche 454. As a result, millions of short reads are produced. Next, these short reads are mapped back to a reference genome or transcriptome. After that, the expression levels are estimated for each gene. Then, the count data are normalized. Finally, statistical testing is adopted to identify differentially expressed genes. Various studies have been shown that the choice of normalization procedure have a decisive effect on identifying differentially expressed genes. The aim of data normalization is to minimize the effects caused by technical variations, such as library size or sequencing depth, gene length, and GC-content. In general, larger sequencing depth results in higher counts, which means that the observed counts are not directly comparable between different samples. Likewise, long genes tend to be mapped larger number of reads. These systematic variations make it difficult to capture true differential expression. Several normalization methods have been developed, such as Total Count, Upper Quality, Median, Trimmed Mean of M values (TMM), Quartile, the Reads Per Kilobase per Million mapped reads (RPKM), and RSEM, to reduce the biases existed in RNA-Seq analysis. However, it is difficult to decide which normalization methods should be used among the various approaches.

In this paper, we evaluated the performance of seven normalization methods in terms of the correlation with qRT-PCR data. The results show that RSEM performs well on various RNA-Seq datasets.

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) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2008-0062611).

Gait identification using Wearable Sensors with Spatio-Temporal Features

Hyun Woo Park, Kyeong Seok Lee, Soo Ho Park, Cheng Hao Jin
Database/Bioinformatics Laboratory, Department of computer science,
Chungbuk National University, South Korea
{hwpark, kslee, soohopark, kimsungho}@dblab.chungbuk.ac.kr

Abstract

In the recent years, biometrics has been very significant to the automatic and user-friendly personal identification based on physiological and behavioral characteristics. Gait identification is recently one of the most important research topics in biometric identification, because gait is unobtrusive and typifies the motion. To acquire a gait dataset for user identification, we used 3L lab's footlogger and this product composed of 8 pressure sensors in insole. The environment conditions we assume that, the number of users is smaller than 5 and all user walk with footlogger everyday. This paper presents methods for footprint-based user identification using gait pattern. We extract various features for user identification and we divided 2 categories domain: time domain and frequency domain.

We defined time domain as a spatiotemporal variation such as stride, swing, and swing interval and swing/stride ratio, stance/stride ratio for both foot that can be extracted from sequence of each step. Stride time is between successive instant of foot floor contact of the same foot, stance is the foot is in contact with the floor, swing is the foot is not in contact with floor. The frequency domain is defined as a temporal variation such as foot angle, accumulated each pressure sensor, heel-strike response first sensor of the footstep and toe-off response last sensor of the footstep.

In our study we used Multilayer perceptorn method to user identification based on the extracted time domain and frequency domain. The results show that, the extracted from time domain and frequency domain features are distinguishable enough to identify the users. In the future work, we consider many people gait data set and compare other classification techniques to identify the users.

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No. 2008-0062611).

Development of System to Measure Length of Moving Object on CCTV Image

Kyu Ik Kim, Sunny Song, Myung-Sic Kim, Jin Suk Kim
Chungbuk National University
han_bando@naver.com

Abstract

CCTV(Closed Circuit Television) have been widely used as the effective approach for preventing crime. In a case of the South Korea's government office, it was first used to prevent the robberies at night in 2002, and it began to expand throughout the whole country. However people have been arguing about the effectiveness of CCTV because of a research that the crime rate decrease slightly for the increase of installed CCTVs. In spite of these problems, Government office, private enterprises and etc have been utilizing to prevent crimes and manage facilities on a small budget.

In this paper, we develop a system to measure length of moving object on CCTV Image. This system can help in decision making about a risk of an intruder to the observer.

1. Introduction

CCTV(Closed Circuit Television) have been widely used as the effective approach for preventing crime. However This way also have strengths and weaknesses. These are strengths. First, a crime rate can be reduced due to CCTV installation. Second, some people will feel safe from crimes when they passed by the location that is installed CCTVs. Third, It is helpful to reduce operational costs for observing about Facility and Security Zone. We could benefit more from adopting CCTV, except for the aforementioned advantages. But there are also the opposite side. First, the proliferation of CCTVs was criticized for violating privacy. Second, people have been arguing about the effectiveness of CCTV because of a research that the crime rate decrease slightly for the increase of installed CCTVs.[1, 2].

In this paper, we develop a system to measure length of moving object on CCTV Image. This system can help in decision making about a risk of an intruder to the observer. This System are not only showing camera images but also helping to decide rapidly through displaying calculation of object's length.

This paper is structured as follows: in section 2 we mention the status of installed CCTVs and problems. Section 3 describe our system. Finally, in section 4 we present our conclusions and future works.

2. The status of installed CCTVs and problems

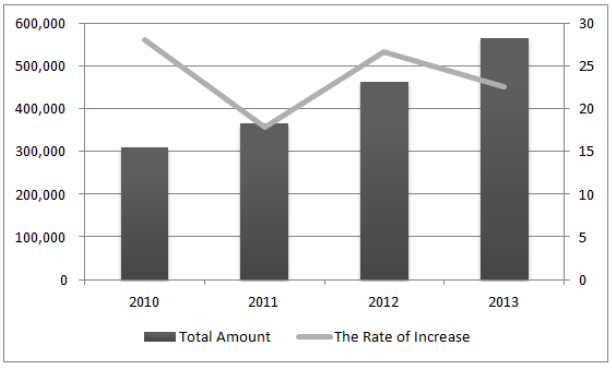


Figure 1. A Statistics of installed CCTVs in government offices

Figure 1. shown the statistics of the installed CCTVs in government offices of the South Korea. An amount of installed CCTVs for facility monitoring and preventing crimes are on an increase trend, However an increasing rate is declining due to problems of violations of privacy[3].

The current CCTV monitoring system is not almost possible to watch on real-time basis, except some place that have an integrated control system and sufficient human resources. Therefore CCTV image have been using to corroborative facts to judge about situations after the crime.

In spite of a important issue to protect the primary school and the kindergarten, the official are hard to place many guards due to lack of budget.

3. Implement of System to Measure Length of Object on CCTV Image

In this paper, our system could contribute to prevent crimes through measuring length of an object and showing on screen. The approach of measuring a length are using length ratio between object in a image.

The restrictions for usage system are defined as follows. First, The CCTV should be fixed. Second, the surveillance zone must be bright. And third, the criterion object have to exist in the zone in order to measure a length.

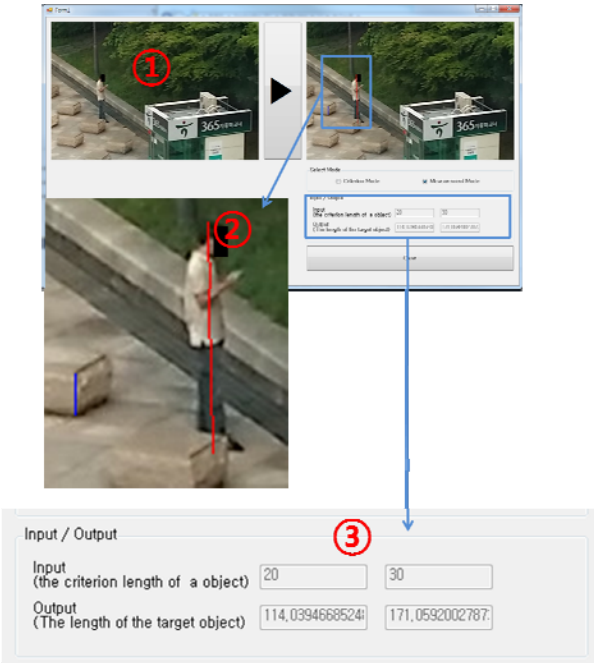


Figure 2. Implement of our System to Measure Length of Object on CCTV Image

Figure 2 present a application program of this system. 1 is shown a image of the current surveillance zone. 2 is a object that we want to measure a length. The blue line is a criterion object. And the red line is a target object. 3 present a measuring length value of object.

4. Conclusion

Our system can display length of object that we want to measure on screen. And it is able to replenish the lack of budget and human resources. We expect to help preventing crimes in such conditions.

In the future work, we need to measure and track length of object on real-time basis.

5. References

[1] W. K. Yang, Y. C. Jeong, "A Study of the Effective Utilization Method on Crime Prevention CCTV - A Focus on University Student's Cognition about Crime Prevention CCTV in University", National Association of Korean Local Government, 15(3), 2013, pp. 102-122.

[2] E. H. Park, J. S. Jeong, "Articles : The effectiveness of CCTV as the crime prevention policy: Using Panel 2SLS Analysis", The Korean Association of Police Science Review, 44, 2014, pp. 39-35,

[3] The statistics of operation about installed CCTVs in a public institution of the South Korea, Statistics Korea, Retrieved from http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=2855, 2014. 06.

Accelerating the morphology operations using a CUDA on Graphics Processing Units

Amartuvshin Renchin-Ochir, Gerelttulga Galaa, Bolormaa Dalanbayar
School of Applied Science and Engineering, National University of Mongolia, Ulaanbaatar, Mongolia
amartuvshin.r@gmail.com

Abstract

Morphology operations has been widely applied in image processing application. Morphology operations most popular method of shape detection and post processing, normally takes a long time to achieve reasonable results, especially for large images. Such performance makes it almost impossible to conduct real-time image processing with sequential algorithms on community computers. Recently, NVIDIA developed CUDA programming paradigm to explore the tremendous computational power for operations on vectors, matrices and high dimensional matrices. In this paper, some morphology algorithms are designed to run on both CPU and GPU computing platforms. Experimental results indicate that the better morphology algorithm on GPUs can achieve up to many times speedup over the version on CPU.

Programmable logic chip based Sinewave generator using Cordic algorhitm

Battogtokh.J, Zorig.B, Bolormaa.D
Department of Electronics and Communication Engineering,
School of Applied Science and Engineering,
National University of Mongolia
Ulaanbaatar, Mongolia
jtogtokh@yahoo.com

Abstract

CORDIC, an acronym for Coordinate Rotation Digital Computer, is a class of shift-add algorithms that rotate a vector in a plane. The CORDIC computing technique a highly efficient method to compute elementary trigonometric functions and this paper presents how to calculate sine and cosine values of the given angle using CORDIC algorithm. Often trigonometric functions are used in embedded applications such as motion control, digital signal processing, filtering and waveform synthesis. Pipeline architectures are used in CORDIC algorithm to reduce the critical path, increases the clock speed. An angle recoding method is used to reduce the latency and obtain the desired angle in least number of iteration. This is implemented using Spartan 3 Fpga XC3S200FT256 and Xilinx ISE design and verification tools.

Technical analysis on 3D reconstruction methods

Erdenebat Budsuren, Baatarbileg Altangerel, Oyun-erdene Namsrai
{free3erkaa, a_bbileg, oyun_erdene79}@yahoo.com

Abstract

Nowadays, use of 3D is increasing in computer graphics, moreover in movie, medical use, and robotics or in everyday needs. In this paper, we demonstrate some basic ideas of 3D reconstruction methods. Reconstruction methods have been developed in many ways. Some of them are used in industrial uses.

1. Introduction

As increasing computing power of technology, we are becoming to create more powerful, creative, and intelligent technology. 3D reconstruction is the process of capturing the shape and appearance of real objects, and building its model in 3D coordinate systems or in virtual environment. In recent times, movie industries, game, medical research needs are why this technology needs to be developed.

Many of 3D reconstruction methods have been developed. But the basic idea of the 3D reconstruction is derived from the human visual perceiving system, or eyes. Every person could classify objects they are seeing by distance from him. This neural function is called brain stereoscopy. Ever since computer was entered into science, expressing human sense in computer have been the one of the most challenging topics.

2. Background

Basic idea of 3D reconstruction is to find depth information from object and create objects in virtual environment using the data. But before researchers can gather depth data they were studying about human perception or computer vision.

Early reconstruction methods include getting main shape from real world. Those kinds of algorithms called Shape from X. The cues getting shape are shape from shape, shape from texture and shape from focus.

3. Methods

3.1. Cues from monocular vision

Those types of methods have similar way to gather data. Data gathered from single view camera.

3.1.2. Shape from shading

One of the monocular cues that human eyes sense depth by their eyes is reaction to brain when we see shaded object or even when we see just shaded drawing. Basic concept of this function is not only based on human brain system, but also on law of reflectance. Shape can be recovered from a single image by human visual system based on the shading information. [1]

The shading algorithms consider on shade reflectance coefficient and reflectance, and light source directions are either known or can be calibrated by the use of a reference object. Under the assumptions of distant light sources and observer, the variation in intensity (irradiance equation) become purely a function of the local surface orientation,

$$I(x; y) = R(p(x; y); q(x; y)); \quad (1)$$

Where (p, q) = (z_x; z_y) are the depth map derivatives and R(p, q) is called the reflectance map. For example, a diffuse (Lambertian) surface has a reflectance map[1] that is the (non-negative) dot product between the surface normal $\check{n} = (p, q, 1)/\sqrt{1 + p^2 + q^2}$ and the light source direction v = (v_x; v_y; v_z),

$$R(p; q) = max\left(0, \rho \frac{pv_x + qv_y + v_z}{\sqrt{1+p^2+q^2}}\right) \quad (2)$$

Where is the surface reflectance factor (albedo).

Using multi-resolution techniques [2] can help accelerate the convergence, while using more sophisticated optimization techniques (Dupuis and Oliensis 1994) [3]can help avoid local minima.

[1] Relates Image Irradiance to surface orientation for given source direction and surface reflectance

In practice, surfaces other than plaster casts are rarely of a single uniform albedo. Shape from shading therefore needs to be combined with some other technique or extended in some way to make it useful. One way to do this is to combine it with stereo matching [4] or known texture (surface patterns) [5]. The stereo and texture components provide information in textured regions, while shape from shading helps fill in the information across uniformly colored regions and also provides finer information about surface shape.

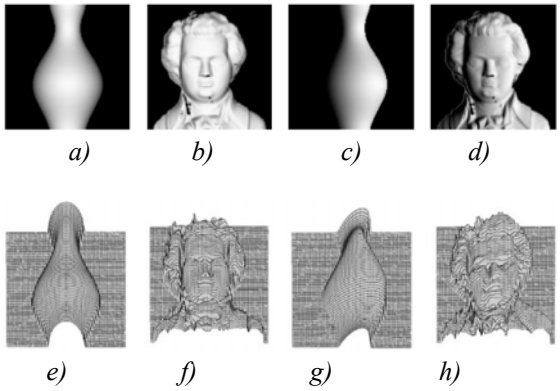


Figure 1. Synthetic shape from shading [20] (a–b) with light from in front (0; 0; 1) and (c–d) with light the front right (1; 0; 1); (e–f) corresponding shape from shading reconstructions using the technique of Tsai and Shah [21]

3.1.2. Photometric Stereo

Photometric stereo is a technique in computer vision for estimating the surface normal of objects by observing that object under different lighting conditions. This method originally introduced by Woodham in 1980 [6].

For each light source, we have a different reflectance map, R1(p, q), R2(p, q), etc. Given the corresponding intensities I1, I2, etc. at a pixel, we can in principle recover both an unknown albedo and a surface orientation estimate (p, q).

For diffuse surfaces (2), if we parameterize the local orientation by n, we get (for non-shadowed pixels) a set of linear equations of the form

$$I_k = p\check{n} * v_k, \qquad (3.3)$$

from which we can recover $p\check{n}$ using linear least squares. These equations are well conditioned as long as the (three or more) vectors v_k are linearly independent, i.e., they are not along the same azimuth (direction away from the viewer).

When surfaces are specular, more than three light directions may be required. In fact, the irradiance equation given in (1) not only requires that the light sources and camera be distant from the surface, it also neglects inter-reflections, which can be a significant source of the shading observed on object surfaces, e.g., the darkening seen inside concave structures such as grooves and crevasses [7].

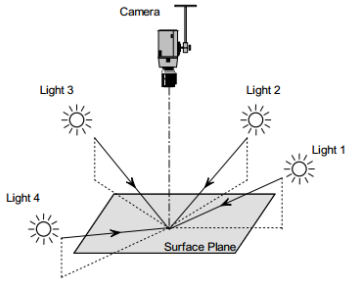


Figure 2. Adjustment of Photometric Stereo

3.1.3 Shape from Texture

The regular repetition of an element or pattern on a surface carries a lot orientation information of this surface. This element is called texel. Shape from texture algorithms require a number of processing steps, including the extraction of repeated patterns or the measurement of local frequencies in order to compute local affine deformations, and a subsequent stage to infer local surface orientation [8].

Recover the orientation of a textured 3D surface from a single image based on the following assumptions

- The image projection is orthographic
- The 3D surface is approximately planar
- The 3D texels are small line segments, called needles
- The needles are distributed evenly in all orientations and evenly on the 3D surface

The deformations induced in a regular pattern when it is viewed in the reflection of a curved mirror, as shown in Figure 3c–d, can be used to recover the shape of the surface [9] [10] [11] [12] [13].

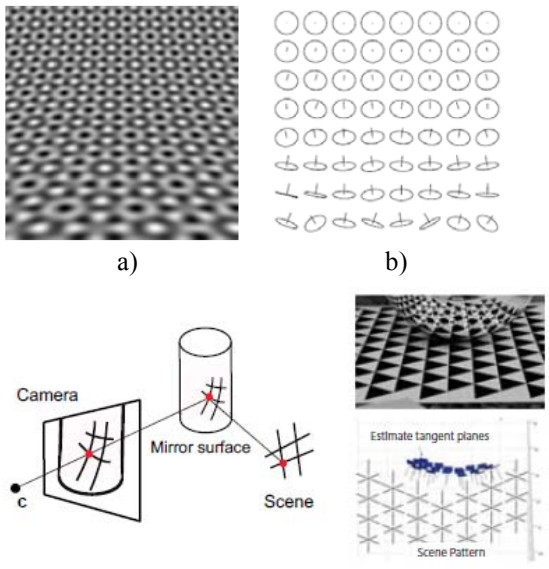


Figure 3. Synthetic shape from texture (a) regular texture wrapped onto a curved surface and (b) the corresponding surface normal estimates. Shape from mirror reflections (c) a regular pattern reflecting off a curved mirror gives rise to (d) curved lines, from which 3D point locations and normals can be inferred.

3.1.4. Shape from Focus

Fundamental to Shape From Focus/Defocus is the relationship between focused and defocused images of a camera. According to paraxial-geometric optics, the basic image formation process is shown in Fig. 1. For an aberration-free convex lens, (i) the radiance at a point in the scene is proportional to the irradiance at its focused image (photometric constraint) i.e., light beams radiated by the object point *P* are intercepted by the lens and are refracted by the lens to converge at a point *P'* on the image plane, and (ii) the position of the point *P* in the scene and the position of its focused image *P'* are related by the well-known Gaussian lens formula (geometric constraint).

- The amount of blur increase in both directions as you moves away from the focus plane. Therefore, it is necessary to use two or more images captured with different focus distance settings [14] [15] or to translate the object in depth and look for the point of maximum sharpness [16]
- The magnification of the object can vary as the focus distance is changed or the object is moved. This can be modeled either explicitly (making correspondence more difficult) or using telecentric optics, which approximate an orthographic camera and require an aperture in front of the lens [15].
- The amount of defocus must be reliably estimated. A simple approach is to average the squared gradient in a region but this suffers from several problems, including the image magnification problem mentioned above. A better solution is to use carefully designed rational filters [17].

Figure 4. (b–c) input video frames from the two cameras along with (d) the corresponding depth map;

3.2. Stereo Imaging

3.2.1. Triangulation

In most 3D reconstruction methods, triangulation is used as basic method to get depth information. Main purpose of triangulation is to determine the location of a point by measuring angles to it from known points at either end of a fixed baseline, rather than measuring distances to the point directly. Triangulations are widely used as a basis for representing geometries and other information appearing in huge variety of applications.

3.3. Model-based reconstruction

In previous methods, we have seen that 3D reconstruction methods that use special requirements to setup environment or some that requires some hardwares to build the reconstructed models. Model based reconstruction methods aims to reconstruct 3D models by using image that could build.

Architectural modeling, especially from aerial photography, has been one of the longest studied problems in both photogrammetry and computer vision [18]. The work by Debevec, Taylor, and Malik [19] solved some image-guided reconstruction tools with model-based stereo matching and view-dependent texture mapping to their system footnotes altogether and include necessary peripheral observations in the text.

Figure 5. Interactive architectural modeling using the Facade system (a) input image with user-drawn edges shown in green; (b) shaded 3D solid model; (c) geometric primitives overlaid onto the input image; (d) final view-dependent, texture-mapped 3D model.

4. Conclusion

Each of techniques gather depth information in their way and each of them have their advantages and disadvantages. Some can be very good used to reconstruct buildings, and some can be good for living things.

Shape from	How many images	Method type	Features
Stereo	2 or more	passive	Works with any objects. But it depends on range from the object
Focus/ Defocus	2 or more	active	Precision of the method not sufficient
Texture	single	passive	Only works with repetition of texels.
Shading	single	passive	Precision of the method not sufficient .
Model-based	1 or more	active	Very useful for architectural model

5. References

[1] B. K. P. a. B. M. J. Horn, "Shape from Shading," *MIT Press, Cambridge,* 1989.

[2] R. Szeliski, " Fast shape from shading.," *CVGIP: Image Understanding,* 53(2), 1991a, pp.129–153

[3] P. a. O. J. Dupuis, " An optimal control formulation and related numerical," 1994.

[4] P. a. L. Y. G. Fua, "Object-centered surface reconstruction: Combining," 1995.

[5] R. a. F. D. White, "Combining cues: Shape from shading and texture.," 2006.

[6] R. J. Woodham, "Analysing images of curved surfaces. Artificial Intelligence,," 1981.

[7] S. K. I. K. a. K. T. Nayar, "Shape from interreflections. International," 1991.

[8] A. Witkin, "Recovering surface shape and orientation from texture.," *Artificial Intelligence,,* 1981.

[9] K. Ikeuchi, "Shape from regular patterns," *Artificial Intelligence,* 22(1), pp. 49–75.

[10] D. a. A. N. Blostein, "Shape from texture: Integrating texture-element extraction," 1987.

[11] J. Garding, "Shape from texture for smooth curved surfaces in perspective projection.," 1992.

[12] J. a. R. R. Malik, " Computing local surface orientation and shape from," 1997.

[13] A. a. F. D. A. Lobay, "Shape from texture without boundaries," *International Journal of Computer Vision,* p. 67(1), 2006, pp. 71–91.

[14] A. P. Pentland, "A new sense for depth of field.," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-9(4), 1987, pp. 523–531.,

[15] S. K. W. M. a. N. M. Nayar, " Real-time focus range sensor," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* p. 18(12), 1996, 1186–1198.

[16] S. K. a. N. Y. Nayar, "Shape from focus.," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 16(8), 1994, pp. 824–831.

[17] M. a. N. S. K. Watanabe, "Rational filters for passive depth from defocus.," *International Journal of Computer Vision,* 27(3), 1998, 203–225.

[18] E. L. a. H. M. Walker, "Geometric reasoning for constructing 3D scene descriptions from images," *Artificial Intelligence,* 37, 1988, pp. 275–290.

[19] P. E. T. C. J. a. M. J. Debevec, "Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach," *In ACM SIGGRAPH 1996 Conference Proceedings,* 1996, pp.11–20.

[20] R. T. P.-S. C. J. E. a. S. M. Zhang, " Shape from shading: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 21(8), 1999, pp.690–706.

[21] P. S. a. S. M. (. Tsai, "Shape from shading using linear approximation.," *Image and Vision Computing,* 12, pp. 487–498.

Session : Keynote Session 3

- Low-Cost Wireless Sensor and Communication System for Landslide Monitoring and Assessment

  Anan Phonphoem

- Combining Tag and Value Similarity for Data Extraction and Alignment

  Weifeng Su

Low-Cost Wireless Sensor and Communication System for Landslide Monitoring and Assessment

Anan Phonphoem
Intelligent Wireless Network Group, Department of Computer Engineering ,
Faculty of Engineering, Kasetsart University, Bangkok, Thailand
anan.p@ku.ac.th

Abstract

Thailand is located in the tropical area with high amount on rainfall for certain months in each year. Areas closed to the mountain with landslide hazard are scattered in the country. Accurate rainfall information are vitally important to the local people. Normally, the rainfall data collectors mostly placed in the convenient places for installation and maintenance such as district or village head office. However, the most effective location for sensing and gathering data should be the place closed to the source of the landslide such as the top and surrounding mountain areas. Also traditionally, to quickly and accurately monitor the situation, human staff and simple tools such as rain gauges are required to operate. Recently, in Krabi province, we have deployed a low-cost wireless sensors along with both short range and long range communication system for landslide hazard assessment. In addition, a debris flow detector enhanced with the wireless communication unit is installed in the area to immediately send alarm once the land slide has been detected. Villagers in the area and landslide researchers can access the current system status and monitoring information via our web-based monitoring system.

Session : Biomedical Informatics

- Practical Problems of Developing Indoor Positioning Systems using Wifi
  Tung Hoang Do Thanh, Tien Nguyen Ba, Binh Ngo Van
- Diagnosing patient with Acute Myocardial Infarction using mRNA Profiles
  Seung Hyeon Yang, Kwang Sun Ryu, Musa Ibrahim M. Ishag
- Exploration of MicroRNA-Based Cancer Classification Using Decision Tree Classifier
  Feifei Li, Minghao Piao, Keun Ho Ryu
- Comparison of Combination of Feature Selection Methods and Classification Methods for Multiclass Cancer Classification from RNA-seq
  Nak Hyeon Choi, Yongjun Piao, Meijing Li, Keun Ho Ryu

Practical Problems of Developing Indoor Positioning Systems using Wifi

Tung Hoang Do Thanh[1], Tien Nguyen Ba[1], Binh Ngo Van[2]
[1]*Information Technology Institute, Academy of Sciences of Vietnam, Vietnam*
tunghdt@ioit.ac.vn , tiendroid@gmail.com
[2]*Hanoi University of Industry, Vietnam*
nvbinh@gmail.com

Abstract

Recently, Indoor LBS is getting interested by many scientists as well as companies because of fast growing up of internet, smart phones, and specially buildings. The systems works well on the ground similar to GPS, but they are replaced the defined GPS trajectories device with signaling equipment installed in a fixed location on the ground. There are many different methods, technologies that can be used for indoor positioning. However, those based on Wifi platform are the most feasible in Vietnam because Access points are set up almost buildings and smart phones are popularly used in Vietnam today. In this paper points out practical problems in developing Indoor positioning systems using Wi-Fi platform, those should be studied in future.

1. Introduction

With more than 770 million smart-phones using a global positioning system, LBS has been gradually changing mobile ecosystem by its versatility and maneuverability. LBS brings a new vitality into mobile advertising, mobile applications, to meet the user's up information and sharing needs, helping the manufacturers in business services, providing trendy revenues advertising, and profitable for retailers as well. However, GPS systems seem uneffect in the home that makes the in inside positioning market become more atractive:

Google has made a determination to upgrade map positioning technology by buying WifiSlam, which invented the indoor positioning technology, with the price of 20 milion USD (according to The Wall Street Journal on March 24th 2013). Moreover Google has applied for indoor positioning technology in some public places, such as airports, shopping centers and stadiums.

However, to restrict opponents in developing positioning technology, Apple and Google have both taken the trade technological barriers. Apple forces their iPhones and iPads products to use location services, - the genuine Apple. Google had retracted the license for using of the two Android phone makers; Samsung and Motorola when the company launched two smart-phones products using the Google's Android operating system, but replace the Google's location services to theirs.

Thus the indoor positioning technology is being developed but concealed, prevented technologically by the "big", the developers also have the inside positioning devices however the devices are difficult to apply in Vietnam, that make us study about the indoor positioning in Vietnam.

Besides, there hardly had any deeply researches in the indoor positioning in VietNam. Most of projects mainly for security area with Bluetooth infrared equipments with foreigner's solutions.

There have been no studies or testing navigation systems using available wifi so far. The first proposed of designing positioning systems in wifi hotspot available buildings, and using, improving the located algorithms to build the test program to the actual results.

1.1 The development ability of positioning technology based to Wifi in Vietnam

In Vietnam, the statistics of smart-phone users is 17 million which are equipped with wireless tuner. The number of internet users is about 40 million and this country is also in the top 20 of internet users in the world. To serve the needs of increasing Internet users and marketing products, the numbers of airing stations are increasing more and more, and even more condense in buildings, commercial centers.

So far, Vietnam is absolutely suitable environment to develop indoor positioning system using wifi. About the market, a practical demand survey is also pointed out the great demand especially in the buildings, commercial centers, stations, airports the numbers of passengers in / out is very frequently leads to the monitoring security need from the administration building while on the customers side, the guidance system need to go to the required position. In hospitals, it is needed to locate the patient / physician especially in the emergency case.

In Section 2 we study and evaluate the indoor positioning technology. Section 3, we propose the algorithm about the located indoor using wifi platform. Section 4 gives conclusion of the paper studies and the future deployment directions.

2. Methods, technologies being used Today

2.1. Technologies

Currently, there are many individuals and organizations working on indoor positioning methods. Every method has its own technology. It is easy to make a list of some common technologies used in indoor positioning process as follows:

•**Infrared**: A number of prototypes developed in research projects are based on infrared signals for realizing positioning methods based on proximity sensing [1]. Unlike radio signals, infrared signals have the advantage in that their emitters have a short range of some meters only and do not penetrate walls. These properties predispose them for use in conjunction with proximity sensing for applications that require to resolve a target's position at least with the granularity of rooms inside a building.

•**Bluetooth**: This technology has higher popularity levels than infrared technology due to slightly stronger wave range. Unregistered 2.4 GHz frequency Bluetooth band is used on the ISM frequency bands [2]. However, cause of the high cost system if it is implemented on a wide area by using the huge numbers of aired stations.

•**Ultrasound** [3]: Ultrasound signals do not penetrate walls and do not require a line of sight between sender and receiver. Unfortunately, their propagation range is very limited, which makes them impracticable for localization in large coverage areas and thus as an alternative to cellular positioning methods. However, inside buildings ultrasound can achieve positioning accuracies in the range of centimeters, assuming that there exists a close-meshed network of transmitters and receivers respectively.

• **WLAN**: WLAN technology has been widely used in some recent years [4]. Typically, when using this technology, the system will measure the signal strength emanating from the access point (AP), the model uses spreading radio waves to identify the distance from the access point to mobile devices or to match them in a data table available. WLAN technology will ensure broad coverage area, have the power to through some obstructions such as walls or devices in the home such as tables, chairs, ...

With the current popularity, the use of WLAN will make the application of navigation systems become easier for the users.

Thus, WLAN technology allows overcoming weaknesses in coverage area of infrared technology and expensive prices of devices using ultrasonic waves. Moreover, WLAN is equipped with a wide range of mobile devices such as laptops or mobile phones.

2.2. Basic Positioning Methods

2.2.1. Proximity Sensing

The easiest and most widespread method to obtain the position of a target relies on the limited range of coverage of radio, infrared, or ultrasound signals [5]. The position of a target is derived from the coordinates of the base station that either receives the pilot signals from a terminal on the uplink or whose pilot signals are received by the terminal on the downlink. In the following discussion, this principle will be denoted asproximity sensing. It is illustrated in Figure 1. Figure 1 (a) shows a configuration with omnidirectional antennas, while figure 1 (b) shows proximity sensing with a sectorized antenna. The known position of the base station that either sends or receives the pilot signals is then simply assumed to be the position of the target.

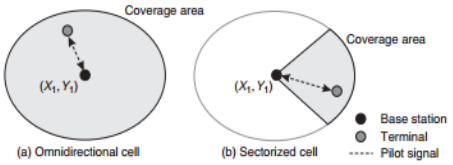


Figure 1. Proximity Sensing

In cellular systems, proximity sensing has become very popular, because it requires only minor modifications on existing infrastructure and causes less overhead. However, the major drawback here is obviously the limited degree of accuracy it provides, which is strongly related with the cell radii, and may vary between 100 m in urban areas and several tens of

kilometers in rural areas. Whether this degree of accuracy is tolerable depends on the respective LBS. In indoor systems, accuracy is much better, which is due to the limited coverage range of the radio, infrared, and ultrasound technologies used here.

2.2.2. Angulation

Angulation is another method to estimate a target's position from the known coordinates of several base stations. In contrast to lateration, the observables here are the angles between the target and a number of base stations. Angulation is also termed Angle of Arrival (AoA) or Direction of Arrival (DoA) [5].

The basic principle behind angulation is illustrated in Figure 2. The angle of an incoming pilot signal is measured at the base station and thus restricts the target's position along a line that intersects both the target's and the base station's position. If the angle to a second base station is taken into account, another line is defined and the intersection of both lines then represents the target's position. Thus, from a theoretical point of view, it is sufficient to make angle measurements at two base stations in order to obtain a position in 2D.

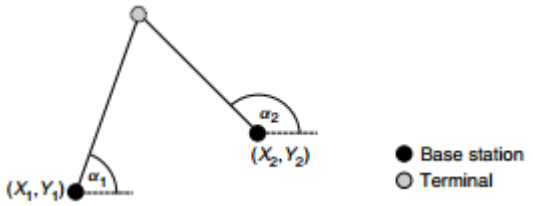


Figure 2. Angulation

However, angulation may suffer from a bad resolution of antenna arrays, and hence the observed angle is rather a rough approximation of the actual angle. This approximation is more accurate if the target is located closer to the base station and vice versa. Furthermore, like for lateration, multipath propagation is a major problem if no line of sight exists between target and base station.

2.2.3. Lateration

For lateration, the observable is either the range or the range difference between the target and a number of at least three base stations [6]. The 2-D solution is then demonstrated in Figure 3. Knowing the range between a terminal and a single base station limits the target's position to a circle around the base station, with the range being the radius of the circle (see Figure 3 (a)). If the range to an additional base station is taken into account, then the target's position can be further reduced to the two points where both circles intersect (see Figure 3 (b)). The range to a third base station then finally leads to an unambiguous position of the target (see Figure 3 (c)).

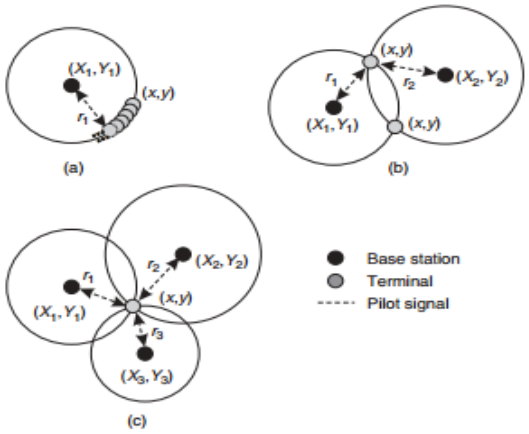


Figure 3. Circular lateration in 2D

The circular multilateration for 3D is shown in Figure 3. Instead of a circle, each range now defines a sphere around each base station where the target may be located. The intersection between two spheres results in a circle and the intersection between three spheres restricts the arget's position to two points. In most cases, one of these points can be canceled, as its coordinates usually represent a rather curious or unrealistic place, for example, somewhere in outer space. Alternatively, the range to a fourth base station can be included to get an unambiguous position. In some systems, for example, GPS, ranges to at least four base stations (i.e., satellites) are needed for clock synchronization.

2.2.4. WLAN Fingerptinting

A location fingerprinting method is often used instead of the radio propagation model, as it can give better estimates of the user's locations for indoor environments [7]. This method is divided into two phases: offline and online phases. During the offline phase, which is also referred to as the training phase, the RSS readings from different APs are collected by the WLAN-integrated mobile device at known positions, which are referred to as the reference points (RPs) to create a fingerprint database, also known as the radio map. Since the orientation of the device's antenna affects the RSS readings, a more comprehensive fingerprint database can be built by collecting RSS readings for different orientations at the

same RP. The actual positioning takes place in the online phase. The mobile device, which is carried by the user collects RSS readings from different APs at an unknown position. Then, these RSS online measurements are compared to the fingerprint database to estimate the user's location by using different methods described in the next section.

Another disadvantage of this fingerprinting approach is the maintenance of such databases. Since the RSS propagation environment varies with time, the accuracy of using the database degenerates over time, as the current RSS readings slowly deviate from the readings in the database.

2.2.5. Dead Rockoning

Dead reckoning is obviously one of the earliest positioning methods that had been applied [8]. It means that the current position of a target can be deduced or extrapolated from the last known position, assuming that the direction of motion and either the velocity of the target or the traveled distance are known. In order to distinguish the position deduced in this way from a position fix (which basically is the last known position derived by an alternative method such as GPS), it is referred to asposition estimatehere. The principle behind dead reckoning demonstrated in Figure 4:

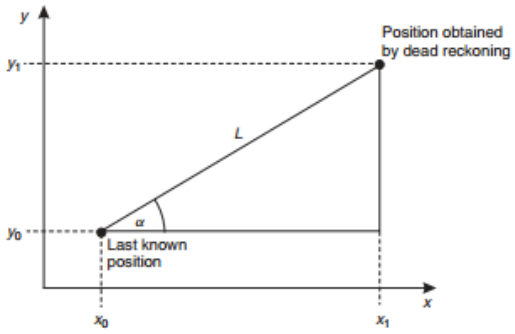


Figure 4. Dead reckoning

Obviously, the crucial point in dead reckoning is to obtain the starting position, the direction of motion as well as the distance and speed respectively. The starting position must be determined by another positioning method, preferably one with increased accuracy such as lateration or angulation, for example, GPS. That is, dead reckoning is not a standalone technique but always used in combination with another method (in the earlier days of navigation, when technologies for positioning were not available, one used well-known control points registered on a map).

2.3 Summary

So there are many methods, indoor positioning technology and every technology, these methods have different strengths and weaknesses. Each technological approach will suit the environment, different platforms. Looking back on wifi platform in Vietnam as well as in the building environment in Vietnam we found the indoor positioning method used in combination with wifi Fingerprinting is more viable if applied in building Vietnam.

3. Proposed using of indoor positioning system available WIFI platform

3.1. System Design

Indoor positioning systems are deployed on cloud platforms which is eveloped on the use of cloud computing resources in the system, the ability to centrally manage and minimize hardware size at the client ... Mobile devices will only interact with the location-based services in the cloud to send and receive data in XML or Json format (Figure 5). All data on maps, points of reference information and location calculations are done on the cloud servers to increase the security of users, support running on other mobile platforms together, management and maintenance becomes much more simple.

Figure 5. Interaction with mobile devices to cloud services

Indoor positioning method based on RSS proposed scheme consists of two stages (Figure 6):

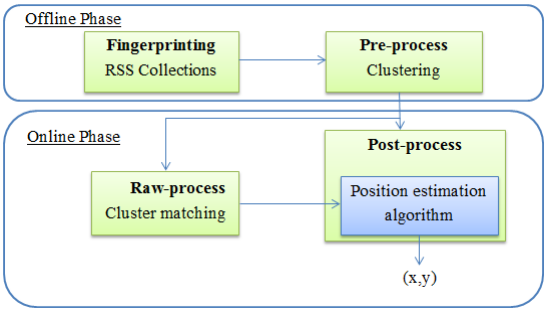


Figure 6. Describe the indoor positioning system based on WiFi availability

Figure 7.a describes the interaction between servers and mobile devices in off-line stage and Figure 7.b describes the interaction between servers and mobile devices in the on-line phase.

Figure 7. Describe the interaction between servers and mobile devices

4.Experimental and evalution

4.1. The test environment

Indoor positioning system was tested on Samsung Galaxy S3 phone at 1 floor of the Institute of Information Technology - Science Academies and Vietnam. Figure 8 shows the test environment with an area of approximately 28m x 22m. The 3 Wi-Fi hotspot are set at 3 positions so that they are capable of covering the entire 5th floor building. All APs support IEEE 802.11n, has coverage in indoor environments of about 35 m2 and are capable of radio signals on both 2.4 GHz and 5 GHz frequency bands.

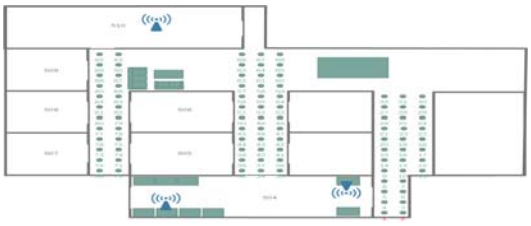


Figure 8. Map of sampling

The deployment of sampling throughout the corridor along with a total of 92 reference points southward and 0.8 m apart each point corresponds to 2 brick size (Figure 8).

Here are the devices and test platform:

•Samsung Galaxy S3: run the Android 4.3 platform, has the task of RSS data collected at every reference point and is used to determine position.

•Server: running Windows Server 2008 R2 Standard 64-bit, 20GB RAM, CPU 2:27 GHz Intel Xeon. WCF services are available to run the request, process and outcome feedback on the phone.

•Applications running Android 4.0 platform includes two main modules that collect data and determine the user's location. Using software tools maps storey building, fitted coordinate system and determine the reference point in the area of interest.

4.2 Test Script

Indoor positioning system is designed based on 2 RSS applications. The first application, also known as Offline applications are used to collect data and transfer data directly to the host (Figure 9a). 2nd application is location-based applications (Online), it uses a database of offline applications to locate objects (Figure 9b).

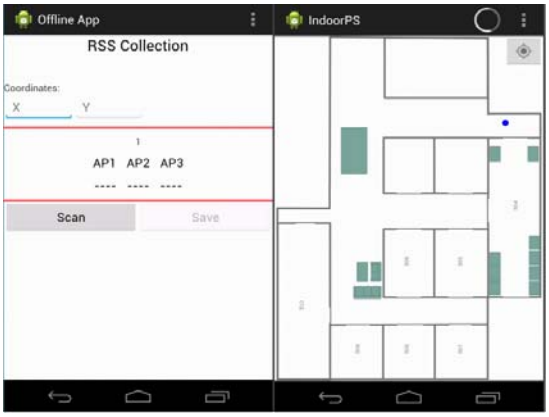


Figure 9. indoor positioning application.

4.3. Offline Application

Before sampling maps revealed a deployment area is attached to the oxygen coordinate system to obtain the coordinates of the reference point. Then, at every reference point users will scan the signal strength from the AP and have enough saved to the data. If the sampling process, users encounter errors in the standing position, direction they can use the delete function and retrieve the data.

4.4. Online Applications

Online applications use databases to be collected from the application offline. The application will automatically turn on WiFi when users open the app; it also automatically scans the signal strength from the AP. This data will be matched with the database to estimate the location. After that, it will display a user's location to the map. However, users can also use the navigation buttons are displayed in the upper right corner of the screen.

4.5. Assessment results

Based on the theory that the ability to estimate the location of the system for high accuracy. However, the actual deployment of emerging issues such as the change of RSS over time, the antenna of a mobile device, location accuracy can be evaluated using the estimated position and the real position is based on the principle of geographic information system of PA Burrough. This method can be used to check the accuracy of object location, space or map drawn on different environments, also known as the Root Mean Square Error (RMSE) values as lower the higher the accuracy.

The change of RSS

During sampling, the RSS values change continuously at the first point of reference, the first direction and a height of about 1.4 m. This is the main cause of error in the estimates. Figure 10 shows the change of RSS value after 30 times of loss scan Samsung Galaxy S3 is set to 10m AP1 and no obstructions.

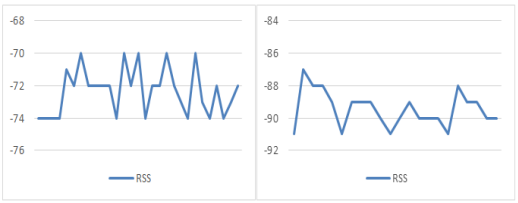


Figure 10: The change of RSS

Figure 10a shows the RSS values change continuously after each scan at the same location and directions ranges from -74 dBm to -70 dBm, the RSS is the difference of 4 units. However, they do have certain stability at -72 dBm signal strength. So the arithmetic mean method can be applied to reduce the position error estimates. For the case of obstruction, RSS value is measured approximately 15m by AP1 and is prevented by two walls (Figure 10b).

In Figure 10b shows the RSS values change in the range -91 dBm to -87 dBm. Thus, deviation RSS is still 4 units and is still the most stable around the average of extremes.

A small cause and effect that RSS is electrical power input per change AP, location receivers in every day only relativity and RSS special values are also affected by mild weather (Figure 11).

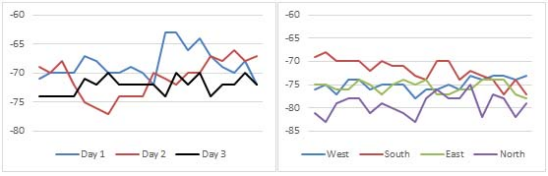


Figure 11. The change of the Day RSS and different

Figure 11a shows significant changes between 3 days with different weather. In which, there is a deviation 2 days of sunshine and a little light on the change similarity, contrast, with drizzling rain, the RSS value vagaries and large deviations RSS (11 units).

Figure 11b shows the change of RSS in 4 different directions. Northern and southern showed the most difference. Mobile device receives in the South has the most intense because of the antenna direction towards mobile devices AP. When turning the device back in the opposite direction (to the north), the signal intensity change in the most significant in comparison with the other direction. Besides, east and west have 2 signal intensity changes in light and average of 2 north and south by the antenna direction of mobile devices nature balance. It is shown that the application can take the average of the four directions will reduce the estimated position error of the system.

The next cause is due Wifi chips are integrated in smart-phones. Currently on the market there are many companies in the phone running the Android operating system. Each company will integrate different wireless chip, and in each product line also integrate different chips. This is a vexing when taken out of the market applications. Figure 11 shows a clear difference in the 3 Samsung, HTC and Google. The most obvious difference is the computer version of the Google Nexus 7 and Asus get the strongest signal intensity compared with the other 2. Thus, the problem of technology for very large errors result oriented and should resolve this later.

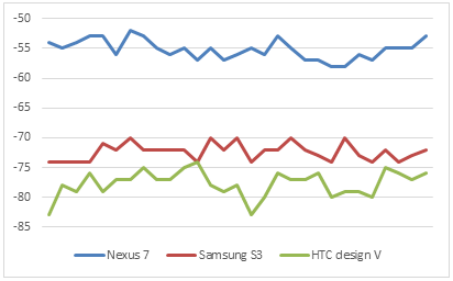


Figure 12. Signal strength of the AP were collected by many different devices

4.6 Accuracy

In navigation stage, the K-NN algorithm gives fairly accurately results when tested on the sampling device (Samsung Galaxy S3). The accuracy location is calculated by RMSE method with k = 3 and k = 10, the larger the RMSE values correspond to greater position errors. During sampling the reference point is taken spaced 0.8m, while implementing the test RMSE = 0.764 m and 3.455m respectively with k = 3 and k = 10. Thus, when k = 3 precision much better when k = 10. This is obviously when the estimated position is calculated as the arithmetic mean k nearest reference point. In Figure 12 shows the position error of 10 attempts at the same location and orientation. Thus the choice of the nearest neighbor k also affects on the estimated position.

Some mobile devices are also included in the testing process on the basis of the Samsung Galaxy S3 data were collected (Figure 13) to test the adaptability of applications running on many different phones.

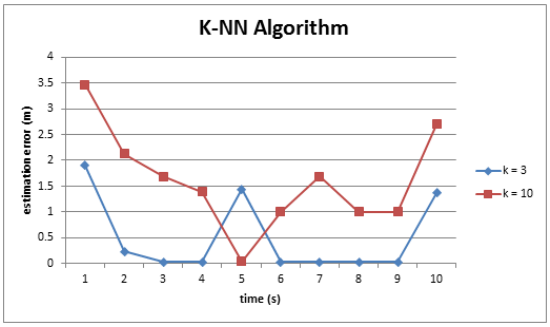


Figure 13. Positioning is based on Euclid distance

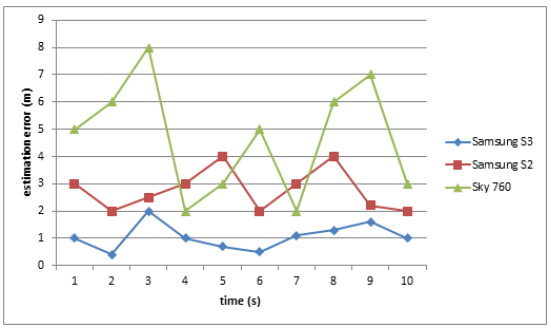


Figure 14. Positioning been tested on many smart-phone devices

Figure 14 shows the Samsung Galaxy S3 has just been tested positioning sample which has the lowest error position one. In contrast, two phones with lower price whose wifi chip works far weaker than the Samsung S3.

Acknowledgement

The author would like to thank IT Academy provides information about the AP test and allows indoor positioning system.

5. References

[1] R. Bruno and F. Delmastro, “Design and analysis of a bluetooth-based indoor localization system,” in Proceedings of IEEE International Conference on Personal Wireless Communications, vol. 27, Venive, Italy, September 2003, pp. 711–725.

[2] N. B. Priyantha, A. K. L. Miu, H. Balakrishnan, and S. J. Teller, “The cricket compass for context-aware mobile applications,” in Proceedings of ACM MOBICOM, Rome, Italy, July 2001, pp. 1–14.

[3] X. Luo, W. J. O’Brien, and C. Julien, “Comparative evaluation of received signalstrength index (RSSI) based indoor localization techniques for construction jobsites,” Advanced Engineering Informatics, vol. 25, April 2011, pp. 355–363.

[4] Mari Saua Svalastot,,“Indoor Positioning Technologies, Services and Architectures”, University of Oslo, Norway , 2007, pp. 18-22.

[5] Hui Liu, “ Servey of Wireless Indoor Positioning Techniques and Systems ”, IEEE transactions on systems, man ,and cybernetics - part C:

applications and reviews, Vol. 37, No.6, 2007, pp. 1067-1071.

[6] Solomon Chan, Gunho Sohn, “Indoor localization using Wifi-based Fingerprinting and Trilateration techniques for LBS applications”, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXVIII-4/C26, Canada, 2012, pp. 2-3.

Diagnosing Patient with Acute Myocardial Infarction using mRNA Profiles

Seung Hyeon Yang, Kwang Sun Ryu, Musa Ibrahim M. Ishag
Database/Bioinformatics Laboratory, Department of Computer Science,
Chungbuk National University, South Korea
{disroway, ksryu, Ibrahim}@dblab.chungbuk.ac.kr

Abstract

In the modern society, cardiovascular diseases rapidly rise because of obesity, smoking, and eating fast-food. Especially, acute myocardial infarction account for most of death caused by cardiovascular diseases. In order to solve this problem, we propose a diagnostic framework of myocardial infarction using RNA expression, which is high dimensional data, following steps: at first, feature selection utilizing differentially expressed gene (DEG), and diagnosing patient with acute myocardial infarction with C4.5 and Naïve Bayes classifier.

1. Introduction

Acute myocardial infarction (AMI) is due to atherosclerotic plaque rupture followed by thrombosis and coronary artery occlusion leading to ischemic myocyte damage so that Myocardium becomes necrosis [1]. This ischemic heart disease remains one of the 10 leading causes of death in the world [2]. To solve this problem, researchers have addressed the problem of identifying biomarkers using gene expression or mRNA Sequence data [3, 4]. In particular Zhang T et al [4] discovered biomarkers responsible for AMI using RNA interaction network, however they didn't attempt to diagnose patients with AMI using those biomarker. Hence, in this paper we present a method for diagnosing patients with AMI, applying the following steps: first, we selected features using differentially expressed gene, shortly DEG, and then diagnose the patient using C4.5 and Naïve Bayes classifier. This paper is organised as follows: Section 2 introduces related work; section 3 describes framework of diagnosing AMI. Experimental results are shown in section 4, while section 5 concludes this paper and presents future work.

2. Related work

Zhang T et al [4] studied how to reveal biomarkers through analyzing interaction network in RNA expression of patients with AMI. They showed that biomarkers may be exhibited by comparing patients of AMI and healthy controls; however they lack diagnosis of patient with acute myocardial infarction utilizing biomarker. Hence, we conducted this study which diagnoses the patient using classifier. Rosalba Giugno et al [5] proposed MIDClass[17] (Microarray Data Classification Method) , which is a kind of associative classification method, utilizing the idea – "gene expression interval values could better discriminate subtypes in the same class". This algorithm follows three steps: first, they filtered insignificant genes using statistical method, and then decided range of interval for gene expression exploiting discretization algorithms, finally extracted association rules from frequent item sets of gene expression interval and to be predicted. Even Though MIDClass shows good accuracy in microarray data, the weakness of this algorithm is that the accuracy changed depends on support of frequent item sets. Table 1 shows an example of the accuracy of this algorithm using Lymphoma Cancer data set published by Rosalba Giugno et al.

Table 1. Accuracy of MIDClass for lymphoma cancer

	threshold:0.1		**threshold:0.05**	
accuracy	86%(ID3,0.1,2)		74%(ID3,0.05,2)	
Confusion matrix	31	1	27	5
	17	9	10	16

Marcin Czajkowski et al [6] applied Multi-Test Decision Tree (MTDT) in order to categorize microarray data. This algorithm attempted to reflect low ranked significant rules. Hence this algorithm allowed several univariate tests in each non-terminal node of the decision tree by fixing C4.5 algorithm. It

showed good performance in mRNA profiles. However, this algorithm is ambiguous to determine the number of test condition and it may easily over-fit training data sets.

3. Framework of Diagnosing AMI

The outline of our proposed framework is illustrated by Figure 1. We extract features from RNA expression data using differentially expressed genes. And then we input data to classifier

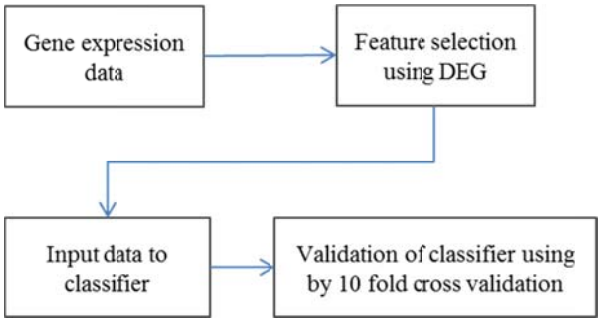


Figure 1. the framework of diagnosing patients with acute myocardial infarction

3.1 Data Set

Same as Zhang T et al [4], we downloaded GSE22229 [7] and GSE29111 [8] data set from Gene Expression Omnibus. GSE29111 [8] is a data set gathered from patient's blood tissue with acute myocardial infarction and unstable angina , that was collected after 7 days and 30 days. GSE22229 includes 12 health control samples. We used amount of 86 samples excluding 18 samples gathered after 30 days from GSE29111 data set.

3.2 Preprocessing: Feature Selection

Characters of RNA expression is not only high dimensional but also few amount of data. Feature selection need to apply data mining techniques because a number of data mining algorithms show good performance in low dimensional data. We utilized differentially expressed genes (DEG) as a feature selection method which is based on t-test and False Discovery Rate, originally this is a list of statistically significant genes [9]. T-test and FDR can be applied following these formulas:

$$\mathrm{t} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (1)[10]$$

$$\mathrm{FDR} = \mathrm{Q_e} = E[Q] = E\left[\frac{V}{V+S}\right] = E\left[\frac{V}{R}\right] \quad (2)[11]$$

We utilize DEG as feature selection method for diagnosis on the basis of biomarkers.

We obtained 45 features from DEG result of 12 patients of AMI and 12 healthy controls. The criterion was $p < 10^{-16}$. Table 2 shows the selected attributes.

Table 2. List of attributes

Probe ID	Gene	Description
1562853_x_at		
200654_at	P4HB	prolyl 4-hydroxylase, beta polypeptide
200707_at	PRKCSH	protein kinase C substrate 80K-H
202358_s_at	SNX19	sorting nexin 19
202513_s_at	PPP2R5D	protein phosphatase 2, regulatory subunit B', delta
202849_x_at	GRK6	G protein-coupled receptor kinase 6
204192_at	CD37	CD37 molecule
205784_x_at	ARVCF	armadillo repeat gene deleted in velocardiofacial syndrome
206892_at	AMHR2	anti-Mullerian hormone receptor, type II
207353_s_at	HMX1	H6 family homeobox 1
208646_at	RPS14P3	ribosomal protein S14 pseudogene 3
208834_x_at	RPL23A	ribosomal protein L23a
209843_s_at	SOX10	SRY (sex determining region Y)-box 10
210295_at	MAGEA10	melanoma antigen family A, 10
210924_at	OLFM1	olfactomedin 1
211445_x_at	NACAP1	nascent-polypeptide-associated complex alpha polypeptide pseudogene 1
211528_x_at	HLA-G	major histocompatibility complex, class I, G
212969_x_at	EML3	echinoderm microtubule associated protein like 3
213356_x_at	HNRNPA1P10//HNRNPA1L2//HNRNPA1	heterogeneous nuclear ribonucleoprotein A1 pseudogene 10///heterogeneous nuclear ribonucleoprotein A1-like 2///heterogeneous nuclear ribonucleoprotein A1
214386_at		
214394_x_at	EEF1D	eukaryotic translation elongation factor 1 delta (guanine nucleotide exchange protein)
216485_s_at	TPSAB1	tryptase alpha/beta 1
216646_at	DSCC1	DNA replication and sister chromatid cohesion 1

217466_x_at	RPS2	ribosomal protein S2
218012_at	TSPYL 2	TSPY-like 2
218378_s_at	PRKRIP 1	PRKR interacting protein 1 (IL11 inducible)
218600_at	LIMD2	LIM domain containing 2
220960_x_at	RPL22	ribosomal protein L22
221379_at		
229320_at	KIAA12 11L	KIAA1211-like
230290_at	SCUBE3	signal peptide, CUB domain, EGF-like 3
230813_at	LEPRE L1	leprecan-like 1
231974_at	KMT2D	lysine (K)-specific methyltransferase 2D
233894_x_at	COL26 A1	collagen, type XXVI, alpha 1
234873_x_at	RPL7A	ribosomal protein L7a
236206_at	FAM53 A	family with sequence similarity 53, member A
236744_at	PHPT1	phosphohistidine phosphatase 1
238150_at		
241486_at		
242735_x_at	ELF2	E74-like factor 2 (ets domain transcription factor)
243917_at	CLIC5	chloride intracellular channel 5
37872_at	JRK	jerky homolog (mouse)
48580_at	CXXC1	CXXC finger protein 1
49327_at	SIRT3	sirtuin 3
61874_at	CACFD 1	calcium channel flower domain containing 1
MI	Labels of	patient with AMI or not

3.3 Classifiers

3.3.1 C4.5 algorithm C4.5 algorithm [12], which is presented by J. Ross Quinlan in 1993, is most popular learning algorithm which is decision-tree based. This algorithm induct learning model using top-down, greedy algorithm manner. It uses information gain ratio as sprit criteria, which normalizes information gain on the basis of information theory. Information gain ratio is calculated as follows:

$$gainRatio(X) = \frac{gain(X)}{splitInfo(X)} \quad (3)\ [12]$$

Where $gain(X)$ calculates the difference between pre-split information entropy and post-split entropy, and $splitInfo(X) = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times log_2\left(\frac{|T_i|}{|T|}\right)$ [12] n is the total number of splits.

3.3.2 Naïve Bayes Classifier [12] Naïve Bayes Classifier or simple Bayesian classifier calculates conditional probability by assuming that attribute set $X = \{X_1, X_2, \ldots, X_d\}$ is independent, given class label Y. This algorithm estimates $P(X_i|Y)$ instead of computing conditional probability of all combination of X.

4. Experimental Results

We performed our experiments using R version 3.1.0 and weka 3.7. R packages used by us are Biobase [14], GEOquery [15], and limma [16] etc. In Weka tool, we utilized J48 (C4.5) and Naïve Bayes classifiers. In order to obtain best classifier for diagnosing patients with AMI, we used 10-fold cross-validation method. Table 3 and Figure 2 show the performance of classifiers.

Table 3. comparison of classifier performance

Items	J48(C4.5)		Naïve Bayes	
Recall	90.7%		81.4%	
Precision	89.9%		92%	
Accuracy	90.698%		81.395%	
Confusion Matrix	6	6	12	0
	2	72	16	58

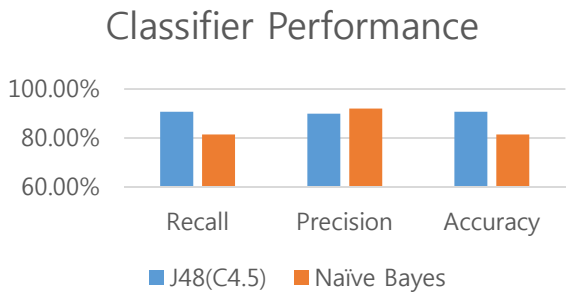


Figure 2. Graphs of classifier performance

Precision is a measure of the accuracy provided that MI class has been predicted. Thus table 4.1 indicate that Naïve Bayes classifier is more exact than J48 (C4.5) classifier. Whereas recall is a measure of the ability of a prediction model to select instances of a certain class from gene expression data set of AMI. Therefore Naïve Bayes is more complete than J48. Accuracy is used for measuring overall performance of classifier. Hence it is evident that J48 show better performance than Naïve Bayes classifier.

5. Conclusion and future work

In this paper, J48 (C4.5) classifier demonstrated good performance to diagnose patients with acute myocardial infarction. We expect that this classifier can be used on clinical field. In the future we will collect more data sets and study data mining algorithms and domain knowledge for improving the model for diagnose patients with acute myocardial infarction.

6. Limitation

In this paper, J48 (C4.5) classifier show better performance than Naïve Bayes classifier due to non-Gaussian distribution caused by the few amount of data volume. Hence experiment need to efficiently verify classifiers using a large scale of universe consist of biomarkers.

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 and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2008-0062611)

7. References

[1] Kristian Thygesen, Joseph S. Alpert, Allan S. Jaffe, Maarten L. Simoons, Bernard R. Chaitman and Harvey D. White, “Third Universal Definition of Myocardial Infarction”, *Journal of the American College of Cardiology*, Vol. 60, 2012, pp. 4-18

[2] www.who.int/mediacentre/factsheets/fs310/en

[3] Michael J et al, “Systematic Review and Collaborative Meta-Analysis to Determine the Incremental Value of Copeptin for Rapid Rule-Out of Acute Myocardial Infarction”, *The American Journal of Cardiology*, Volume 113, Issue 9, 2014, pp. 1581–1591

[4] Zhang T, Zhao LL, Zhang ZR, Fu PD, Su ZD, Qi LC, Li XQ, Dong YM., “Interaction network analysis revealed biomarkers in myocardial infarction”, *Molecular Biology Reports ISSN: 0301-4851 (Print) 1573-4978 (Online)*, 2014

[5] Rosalba Giugno, Alfredo Pulvirenti,Luciano Cascione, Giuseppe Pigola,Alfredo Ferro, “MIDClass: Microarray Data Classification by Association Rules and Gene Expression Intervals”, *PLOS ONE www.plosone.org*, Volume 8, 2013

[6] Marcin Czajkowski, Marek Grze´s, Marek Kretowski, “Multi-test decision tree and its application to microarray data classification”, *Artificial Intelligence in Medicine*, 61, 2014 , pp. 35–44

[7] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE22229

[8] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29111

[9] Ju Han Kim, “Genome data analysis”, *Beommun education co., ltd*, 2012, pp. 40-47

[10] http://en.wikipedia.org/wiki/Student's_t-test

[11] Benjamini, Yoav, Hochberg, Yosef , "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society*, Series B 57 (1), 1995, pp. 289–300.

[12] J. Ross Quinlan, “C4.5: programs for machine learning”, *Morgan Kaufmann Publishers Inc. San Francisco, ISBN: 1-55860-238-0*, 1993, pp. 17-26

[13] Pang-ning Tan, “Introduction to data mining, chapter8, cluster analysis: Basic Concepts and Algorithms, Pearson Education”, 2006, pp. 227–240

[14] V. J.Carey, D. M. Bates, B.Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge et al, “Bioconductor: Open software development for computational biology and bioinformatics R. Gentleman”, *Genome Biology*, Vol. 5, R80, 2004

[15] Davis, S. and Meltzer, P. S., “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”, *Bioinformatics*, 14, 2007, pp. 1846-1847.

[16] Smyth, GK Limma, “linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor”, *Springer*, 2005, pp. 397-420.

[17] http://ferrolab.dmi.unict.it/MIDClass.html

Exploration of MicroRNA-Based Cancer Classification Using Decision Tree Classifier

Feifei Li, Minghao Piao, Keun Ho Ryu
Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea
{feifeili, bluemhp, khryu}@dblab.chungbuk.ac.kr

Abstract

The researches about microRNA expression profiles showed a new aspect of human cancer classification. Since the high dimensionality of the miRNA expression data, the feature selection is necessary to reduce the dimensionality. However, most of the feature selection algorithms used for miRNA expression data have a drawback that they just consider the high-ranking features whereas remove the low-ranking features. But the low-ranking features may perform will in terms of classification accuracy since one miRNA may have influence for more than one kind of cancer. Considering this shortcoming, we proposed the information gain feature selection method to select both the high and low-ranking features, and then used the decision tree classifier to construct the multi-class human cancer classification. For decision tree classifier, we adopted different algorithms, including C&R tree, CHAID method, exhaustive CHAID method, and C5.0 algorithm. After numerous tests, the results proved the usefulness of the m-to-n feature subset in cancer classification since considering the low-ranking miRNAs can get higher classification accuracy.

1. Introduction

The first microRNA (miRNA) was characterized in early 1990s. However, the miRNAs were not recognized as a distinct class of biological regulators with conserved functions until early 2000s. With more and more miRNAs have been discovered, it has been proved that the dysregulation of miRNAs has association with disease, especially human cancer [1, 2]. Chronic lymphocytic leukemia [3] is the first discovered human disease which is known to be associated with miRNA deregulation. Since then, many miRNAs have been found to have links to some types of human cancers.

With these discoveries, many analyses about cancer classification using miRNA expression profiles have been done by using different machine learning methods since the miRNAs have the significant advantage in cancer classification compared with the messenger RNAs (mRNAs). Unlike measurements of mRNA, which must be translated to protein to have a biological effect, miRNA expression levels represent more closely the functional level of the gene. And it has also been proven that some tumor samples that are difficult to classify with mRNA expression profiles can be discriminated with high accuracy by using miRNA profiles.

Although there are many studies using different feature selection methods to do the cancer classification based on miRNA expression profiles, there still has a lot of room for improvement. The high dimensionality of the miRNA expression profiles lead to the necessary of the feature selection process. However, for feature selection, these methods just consider the condition that the relationship between feature and class is 1:1 or n:1, but not consider the condition that the relationship between feature and class is 1:n or m:n. But since the miRNA expression data is a special kind of data, one miRNA may have influence to more than one type of cancers. If using the traditional feature selection algorithms, these miRNAs may be deleted, since they will be considered as the low-ranking features. But this kind of miRNAs are also very important, removing them may lead to the loss of important information. Therefore, we made a new hypothesis that consider both of the high and low-ranking features to cover all the cases (1:1, n:1, 1:n, m:n) can get better accuracy in the cancer classification.

In this work, we chose the information gain feature selection method to build the m-to-n feature subset. Also we adopted the decision tree classifier to construct the multi-class classification using miRNA

expression profiles. Different algorithms were adopted to build the decision tree classifier including the classification and regression (C&R) tree, the chi-squared automatic interaction detection (CHAID) method, the exhaustive CHAID method, and the C5.0 algorithm. After numerous tests, the results proved the usefulness of the m-to-n features in cancer classification since the results showed that considering both the high and low-ranking miRNAs can get higher classification accuracy than just considering the high-ranking miRNAs.

The remainder paper is organized as follows. The methods used in this work are introduced in Section 2. The data and performance evaluation are discussed in Section 3. The conclusion of our work is presented in Section 4.

2. Classification Method

The decision tree classifier [4] is a structure in the form of a tree. In the tree, each internal node represents a test on an attribute each branch represents the outcome of the test and each leaf node represents a class label. The path from the root to the leaf represents classification rules.

The classification and regression (C&R) tree [5] is a recursive partitioning method that first exams the input fields to find the best split, then the split defines two subgroups, each of which is subsequently split into two more subgroups, this step is repeated until one of the stopping criteria is met. The chi-squared automatic interaction detection (CHAID) method [6] builds decision trees by using chi-square statistics to identify optimal splits. The exhaustive CHAID method is a modification of the CHAID method that does a more thorough lob of testing all possible splits and thus takes more time to compute. The C5.0 algorithm can be used to build either a decision tree or a rule set, and it splits the samples based on the maximum information gain. A decision tree is just a direct expression of the C5.0 algorithm. A rule set is a set of rules that aims to make predictions for individual records, and these rules sets are derived from the decision tree using the C5.0 algorithm.

3. Experimental Results

We built the multi-class classification using miRNA expression profiles from the work of Lu et al [7]. The data set consists of 73 tumor samples, these samples belong to five types of cancer including ten colon samples, nine pancreas samples, ten uterus samples, twenty-six B-cell acute lymphoblastic leukemia samples, and eighteen T-cell acute lymphoblastic leukemia samples. Each sample has expression value of 217 miRNAs. The structure of the data set is shown in Figure 1.

Then we used the information gain method to do the feature selection process, and ten high-ranking features and seventeen low-ranking features were selected to build the m-to-n feature subset. Then we used these features to construct cancer classification using decision tree classifier. We compared the accuracy of just using the m-to-n feature subset with the accuracy of using the ten high-ranking features. The result is shown in table 1. When considering low-ranking features the CHAID and the exhaustive CHAID training method can get higher accuracy than just using the high-ranking features. While for C&R Tree and C5.0 training method they get the same results for both using m-to-n feature subset and using just high-ranking features.

EAM340	EAM341	EAM346	EAM352	EAM361	Name
6,3339	9,04258	5,6828	7,1504	5,5897	T_COLON
5,8201	9,056	5,5491	7,2669	5	T_COLON
5,8178	9,07359	5,8547	7,2429	5	T_COLON
5,9502	9,54018	5,4945	5,9427	5	T_COLON
6,5343	10,134	6,5277	6,811	5	T_COLON
6,4842	10,6124	6,2339	8,0337	5	T_COLON
6,1697	9,76274	6,5062	7,8925	5	T_COLON
5,8322	8,42992	5,2409	6,1969	5	T_COLON
6,3231	9,91142	6,3032	7,0883	5	T_COLON
6,0302	9,05413	6,2983	7,3683	5	T_COLON
6,7509	9,22531	6,4543	8,2904	6,1595	T_PAN
5,7524	9,17652	5,955	7,6729	5,4475	T_PAN
5,9324	8,22399	5	7,1674	5,238	T_PAN
6,4701	9,44971	6,7535	8,1606	6,0499	T_PAN
6,6638	9,6194	6,8613	8,0383	6,2913	T_PAN
5,9516	7,96842	5	7,2511	5	T_PAN
5,9572	8,45076	5,6635	7,6385	5	T_PAN
6,3874	8,69437	6,0681	7,9794	5	T_PAN
5,8533	8,8288	5,4102	7,9077	5	T_PAN
6,6404	8,61392	5,8023	7,7623	5,4454	T_UT
6,9255	10,1638	6,8346	8,0155	6,0959	T_UT

Figure 1. The miRNA expression profiles using in this data set

4. Conclusion

In this paper, we used the decision tree classifier to build multi-class cancer classification, and built the m-to-n feature subset considering both high and low-ranking miRNAs. For decision tree classifier, we chose six different kinds of training methods including the C&R tree method, the CHAID method, the exhaustive CHAID method, the C5.0 for decision tree method, and the C5.0 for rule set method. After numerous tests, we found that considering both the high and low-ranking features can get higher accuracy than just considering high-ranking method when using CHAID and exhaustive CHAID method.

Table 1. The classification accuracy (%) of decision tree classifier with different training methods

Relationship	C&R Tree	CHAID	E-CHAID	C5.0-Decision Tree	C5.0-Rule Set
m:n	98.63	97.26	98.63	95.89	95.89
1:1 or n:1	98.63	91.78	90.41	95.89	95.89

1:1, n:1 and m:n indicate the relationship between feature and class: 1:1 and n:1 mean high-ranking features; m:n means both high and low-ranking features.

Our work indicated the usefulness of the low-ranking miRNAs. But the right choice of training method for cancer classification is also very important. For decision tree classifier, the CHIAD method and exhaustive CHAID method are right choices for low-ranking miRNA expression data, while other three methods used in this work are not suitable for low-ranking miRNAs. There is also another problem in our work that it is difficult to select the right number of low-ranking features. So in the future work, we will build the optimal algorithm to select the low-ranking miRNAs automatically, and find the right classification method to construct cancer classification using the m-to-n feature subset.

Acknowledgements

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA (National IT Industry Promotion Agency).

5. Reference

[1] He, L., J. M. Thomson, M. T. Hemann, E. Hernando-Monge, D. Mu, S. Goodson, S. Powers, C. Cordon-Cardo, S. W. Lowe, G. J. Hannon, and S. M. Hammond, "A MicroRNA Polycistron As A Potential Human Oncogene," *Nature*, 435(9), 2005, 828-833.

[2] Mraz, M., S. Pospisilova, K. Malinova, I. Slapak, and J. Mayer, "MicroRNAs in Chronic Lymphocytic Leukemia Pathogenesis and Disease Subtypes", *Leuk Lymphoma,* 50(3), 2009, 506-509.

[3] Mraz, M. and S. Pospisilova, "MicroRNAs in Chronic Lymphocytic Leukemia: From Causality to Associations and Back", *Expert Review of Hematology*, 5(6), 2012, 579-581.

[4] Mishra, A. K. and H. Chandrasekharan, "Analysis and Classification of Plant MicroRNAs Using Decision Tree Based Approach", *Advances in Intelligent Systems and Computing*, 249, 2014, 105-114.

[5] Emrouznejad, A. and A. L. Anouze, "Data Envelopment Analysis with Classification and Regression Tree - A Case of Banking Efficiency", *Expert Systems*, 27(4), 2010, 231-246.

[6] Belaid, A., T. Moinel, and Y. Rangoni, "Improved CHAID Algorithm for Document Structure Modelling", *SPIE Proceeding*, 7534, 2010, 1-10.

[7] Lu, J., G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub, "MicroRNA Expression Profiles Classify Human Cancers," *Nature*, 435, 2005, 834-838.

Comparison of combination of feature selection methods and classification methods for multiclass cancer classification from RNA-seq gene expression data

Nak Hyeon Choi, Yongjun Piao, Meijing Li, Keun Ho Ryu
Database/Bioinformatics Laboratory, Department of Computer Science,
Chungbuk National University, South Korea
{nak, pyz, mjlee, khryu}@dblab.chungbuk.ac.kr

Abstract

RNA-seq is recently developed to solve biological questions by sequencing technology. RNA-seq data make digital gene expression data by some protocol like sequencing, alignment, and estimating gene expression. TCGA site support well-made RNA-seq gene expression data from a lot of many kinds of cancer patients. We use these data for multiclass cancer classification. For exact comparisons with each sample, TMM-normalized genes expression data are used and combinations of feature selection and classification method apply to this data. Combinations of SU & NB, CFS & SMO show high performances than others..

1. Introduction

Microarray was a good tool for researching gene expression studies about cancer diagnosis and multiclass cancer classification. But this technology had one limitation related to probes design, we need to know about target information about DNA sequences or RNA sequences. As this reason gene expression studies was prone to difficult for finding new gene expression information.

RNA-seq is a new tool for researching deeply DNA or RNA. Gene expression data can be obtained from RNA-seq reads data. RNA-seq extract mRNA from a target cell or tissue and make a lot of reads. The number of reads count are used for estimating gene expression abundances after alignment process which specifies reads position to reference genome. RNA-seq don't need probe design like the microarray technology to know gene expression. That RNA-seq characteristic make new gene expression to be found less difficultly than the microarray technology.

But RNA-seq data generate different the number of total read counts from each sample. It mean that gene expression from measuring RNA-seq reads data should be normalized to reduce technical bias for acute comparison with each samples. Two normalization methods, which are DESeq and TMM, are evaluated as good performance (Marie-Agnes et al, 2012). These methods reduce technical bias related to sequencing depth.

RNA-seq gene expression data are similar with microarray gene expression data. A lot of features are more than the number of samples. One column represent one sample and one row is gene ID and gene expression value. But format of raw gene expression data is different. Microarray gene expression raw values are continuous due to measuring dye-light densities but RNA-seq gene raw values are discrete due to measuring reads

Although RNA-seq gene expression data have huge multi-dimension, All features are unnecessary for multiclass cancers classification. Because all features are not considered as cancer genes and some features can distort classification results. Therefore feature selection technique is important role for good classification performance. Feature selection technique reduce redundant attributes, unrelated features and give high score to good features according to their algorithm and then remain a features set.

In this paper, we prove that RNA-seq gene expression data are appropriate for multiclass cancer classification and use feature selection technique to improve multiclass cancer classification performance

.

2. Material & Methods

2.1. Data sets

RNA-seq data can obtain from NCBI web sites and TCGA web sites. TCGA web site support many broad type of cancers and many kinds of formats of cancers data. We downloaded RNASeqv2 data which is generated from state of the art sequencer named as IlluminaHiSeq.

We downloaded each thirty samples of breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), liver hepatocellular carcinoma (LIHC), lung adeno-carcinoma (LUAD), ovarian serous cystadeno-carcinoma (OV) prostate adenocarcinoma (PRAD), thyroid carcinoma (THCA). Downloaded cancer data contents include 20531 of gene id, raw count for a gene, scaled estimation, transcription id. We only use gene id and raw count and then make gene expression table for a multiclass cancer classification experiment. We use only each thirty samples because it is too difficult to acquire over thirty patients data of each cancer from RNA-seq in general

2.1. Data sets

Each sample of multiclass cancer gene expression have the different total number of reads, hence direct comparison of each sample is not available. TMM normalization reduces bias of sequencing depth. Each gene expression of each sample is normalized by TMM.

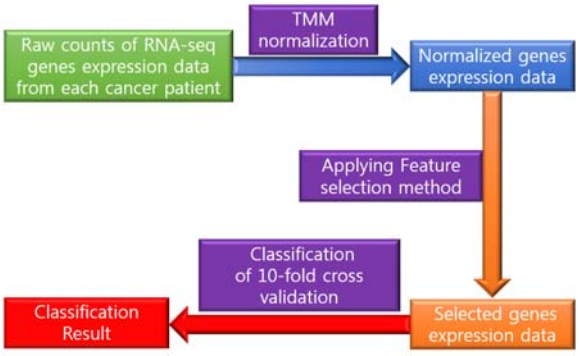


Figure 1. Flow diagram of experiment

Figure 1 shows our flow of experiment. After raw counts of gene expression in gene expression table are normalized, various feature selection methods are applied to improve classification performance. Feature selection methods are symmetrical uncertainty (SU), ReliefF, correlation based feature selection (CFS), CV, gain ratio (GR), chi-squared (Chi), Significance. 6 feature selection methods are used by WEKA. After. 100 features are selected from 20531 features by feature selection methods due to exact comparison of effect of each feature selection method. After normalized gene expression data are through 7 feature selection method, 100 feature of genes expression data are classified by 5 kinds of classifier which are naivebayes(NB), sequential minimal optimization algorithm for training support vector machines(SMO), nearest-neighbor-like algorithm using non-nested generalized exemplars(NNge), J48, logistic model trees(LMT) by WEKA

3. Results and Discussion

After 7 feature selection methods and 5 classification methods are applied to normalized genes expression data, Table 1 show multiclass cancer classification accuracy result.

Table 1. Multiclass cancer classification accuracy results of 10 fold cross validation

Feature Selection	Selected genes	NB	SMO	NNge	J48	LMT	Average
None	20531	**97.0833**	96.6667	95.4167	89.5833	93.75	94.5
SU	100	**98.75**	97.9167	**97.5**	93.75	97.0833	**97**
ReliefF	100	96.25	97.5	96.25	95	97.0833	96.41666
CFS	100	97.5	**98.3333**	95	92.5	**97.9167**	96.25
CV	100	48.3333	46.6667	55	78.3333	79.5833	61.58332
GR	100	95.8333	93.3333	91.25	92.0833	93.75	93.24998
Chi	100	98.75	95.8333	95.8333	95	95.4167	96.16666
Significance	100	98.3333	96.6667	95.4167	**94.1667**	94.5833	95.83334

From the classification accuracy table, NB is the highest than other classification methods in none feature selection methods & all genes. TCGA data are very well made, therefore classification results are very high, but classification of all genes spend many cost.

After feature selection methods are applied, many classification performances are increased. In NB, SU feature selection method show the highest accuracy. SU feature selection method improve performances of other classification methods. But CV feature selection method show bad performances of classification methods. Hence CV is not appropriate for RNA-seq gene expression data

We can show that SU and CFS feature selection methods be able to select good features sets than others. Hence classification performances of SU and CFS are better than other feature selection methods

4. Conclusion

From results of our experiment, Combinations of SU & NB and CFS & SMO show good accuracy from RNA-seq genes expression data than other

combinations. But CV feature selection methods are bad permances for RNA-seq genes expression.

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency)

5. References

[1] Carlos J, Q. Isaac Moro-Sancho, "Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods", *Expert Systems with Applications*, ELSEVIER, vol.39, Issue8, 2012, pp. 7270-7280.

[2] G. Victo Sudha George, V. Cyril Raj, "Review on Feature Selection Techniques and the impact of SVM for Cancer Classification using Gene Expression Profile", *International Journal of Computer Science & Engineering Survey(IJCSES), AIRCC*, vol.2. No3, 2011.

[3] Marie Agnes Dillies et al, "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis", *Briefings in Bioinformatics*, vol.14 Issue.6, 2013, pp. 671-683.

[4] Moulos P, Kanaris I, Bontempi G, "Stability of feature selection algorithms for classification in high-throughput genomics data sets", *Bioinformatics and Bioenineering (BIBE*), 2013 IEEE 13th international Conference on, vol., 2013, pp. 1-4.

[5] Jakob Loven et al, " Revisiting Global Gene Expression Analysis", *Cell, ELSEVEIR*, vol.151, Issue 3, 2012, pp. 476-482.

[6] Xiaobo Li et al, "Comparison of feature selection methods for cancer classification based on microarray data", *Biomedical Engineering and Informatics (BMEI*), *2011 4th International Conference on*, vol.3, 2011, pp.1692-1696.

[7] Yvan Saeys, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics", *Bioinformatics, OXFORD JOURNALS*, vol.23, Issue.19, 2007, pp. 2507-2517

Session : PSM Session

- Shape Representation Using Morphological Granulometries
  Nipon Theera-Umpon
- A Personalized u-commerce Recommender System using Bayesian learning and Weighted Preference
  Seon-Phil Sunny Jeong
- Distance Metric Learning for Face Recognition
  Suvdaa Batsuuri
- Design for Triple Helix model framework using information of bibliography
  Gwi Suk Gim, Ho Sun Shon, Byung Jun Cho, HyungChul Rah, So Young Kim
- Big Data based Framework Design for Korean Patients with Acute Myocardial Infarction
  Changwoo Woo, Wooyeong Jang, Ho Sun Shon, Eung-Do Kim, Gilwon Kang

Shape Representation Using Morphological Granulometries

Nipon Theera-Umpon
Senior Member, IEEE, Biomedical Engineering Center, Electrical Engineering Department, Faculty of Engineering, Chiang Mai University, Chiang Mai, Thailand
nipon@ieee.org

Abstract

In this special talk, an application of morphological granulometries in shape representation of white blood cell nuclei is reviewed. Morphological granulometries are one of powerful tools for objects' shape representation. It is derived from successive applications of morphological opening which result in the granulometric size distributions or pattern spectra. The pattern spectra of different types of white blood cells nuclei are illustrated. The spectra show that morphological granulometries can be used to represent nuclei shapes of white blood cells. A simple set of features extracted from the pattern spectra can be used efficiently for white blood cell classification.

A Personalized u-commerce Recommender System using Bayesian learning and Weighted Preference

Young Sung Cho[1], Keun Ho Ryu[1] In-Bae Oh[2], Seon-phil Jeong[3]
[1]*Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*
youngscho@empal.com, khryu@dblab.chungbuk.ac.kr
[2]*Chungbuk Health & Science University, Chungbuk, Korea*
iboh@chsu.ac.kr
[3]*DST, BNU-HKBU United International College*
spjeong@uic.edu.hk

Abstract

This paper proposes a personalized u-commerce recommender system using Bayesian learning and weighted preference under ubiquitous computing environment which is required by real time accessibility and agility. In this paper, using an implicit method without onerous question and answer to the users, it is necessary for us to keep the Bayesian learning and the FRAT(Frequency, Recency, Amount and Type of merchandise or service) score, in order to calculate the weight based on the number of customers by the rank of customer's score, to reflect frequently changing preference of customer as time goes by emphasizing the important items and to improve the accuracy of recommendation with high purchasability. To verify improved performance of proposing system, we make experiments with dataset collected in a cosmetic internet shopping mall.

Keywords: *FRAT; Bayesian network; Clustering; k-means algorithm*

1. Introduction

Due to the advent of ubiquitous computing environment, it is becoming a part of our common life style that the demands for enjoying the wireless internet using intelligent portable device such as smart phone, PDA and smart pad are increasing anytime or anyplace without any restriction of time and place[2][4]. In these trends, the personalization becomes a very important technology which could find exact information to present users. The recommendation system helps customers to find items easily and helps the ecommerce companies to set easily their target customer by automated recommending process. Therefore, customers and companies can take some benefit from recommendation system. The possession of intelligent recommendation system is becoming the company's business strategy. A recommendation system using segmentation analysis technique to meet the needs of customers, it has been actually processed the research[1-4]. We can improve the accuracy of recommendation using FRAT method for item segmentation and clustering by item category so as to be able to reflect the attributes of items. As a result of that, we can propose a personalized u-commerce recommender system using Bayesian learning and weighted preference. The next section briefly reviews the literature related to studies. The section 3 is described a new method for recommendation system in detail, such as system architecture with sub modules, the procedure of processing the recommendation, the algorithm for proposing system. The 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. FRAT

The RFM(Recency, Frequency and Monetary) score is correlated to the interest of e-commerce[2]. It

is necessary for us to keep the analysis of RFM to be able to reflect the attributes of the item in order to find the items with high purchasability. For the analysis of FRAT, RFM formula was further expanded by Robert Kestnbaum who added a new factor known as T for type of merchandise or service purchased and introduced the new formula identified by the acronym FRAT where F stands for frequency, R for recency, and A for amount. The FRAT score will be shown how to determine the customer as follows, will be used in this paper. The categories (F, R, A, T) have five bins. The variables (A, B, C, D) are weights.

$$\text{FRAT score} = F \times A + R \times B + A \times C + T \times D \quad (1)$$

He suggested that what a person buys at present would be indicative of what that person would buy in the future. In this paper, we can make the task of preprocessing for clustering purchase data to join customer's data using demographic data and FRAT scoring method to recommend the item they really want exactly. In this paper, we can use the analysis of FRAT in order to consider the importance of type of many different items in purchase data, then reflect their different importance of type of merchandise and adjust the preference of weight for recommending service by emphasizing the important transactions. And also, we can use frequent pattern mining for recommending service so as to meet the needs of customers considering type of merchandise or service on e-commerce.

2.2. Bayesian Network

Bayesian networks can be used to model the joint probability distribution of multiple random variables. With the Bayesian network, we formulate a item preference model in the form of a joint probability distribution. In the case of item recommendation, the problem is finding items that a given user is likely to rate highly. For this purpose, we calculate the conditional probability for the target user U=u, the candidate item C=c and then recommend items in order of probability.

Alternatively, we may calculate the conditional probability for the target user and rating to find items that are highly likely to obtain a positive rating. The recommendation system may receive user feedback for final purchase behavior, and periodically, the system updates the parameters of the item preference. Bayesian network model using final purchase data by using the Bayesian inference engine as the decision of behavior of buying additional item in order to increase the precision of the recommendation. Although the preference model can be used in many ways, here, we explain the typical ways for item recommendation. Here, since a recommendation system can use the same item preference Bayesian network model can have two type of the calculation of probability, one is prior probability, the other is posterior probability. The users can be commonly used to update the parameters of the model and thus increase the precision of both the recommendation and promotion. Bayesian probability measures a degree of belief. Bayes theorem then links the degree of belief in a proposition before and after accounting for evidence. For proposition C_i and evidence X,

· $P(C_i)$, the prior, is the initial degree of belief in C_i

· $P(C_i|X)$, the posterior, is the degree of belief having counted for X.

· the quotient $P(X|C_i)/P(X)$ represents the support X provides for C_i

Bayes' theorem gives the relationship between the probabilities of C_i and X, $P(C_i)$ and P(X), and the conditional probabilities of C_i given X and X given C_i, $P(C_i|X)$ and $P(X|C_i)$. For example, suppose an experiment is performed many times. $P(C_i)$ is the proportion of outcomes with property C_i, and $P(X)$ that with property X. $P(X|C_i)$ is the proportion of outcomes with property X *out of* outcomes with property C_i, and $P(C_i|X)$ the proportion of those with C_i out of those with X. In Bayesian inference, the posterior distribution is proportional to the product of the likelihood and the prior distribution. For parameters C_i and data X. It is most common form as follows[5]. For events C_i and X, provided that $P(X) \neq 0$,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}, \quad 1 \leq i \leq m \quad (2)$$

The denominator is the marginal likelihood of the data, which is the integral of the likelihood against the prior distribution. In many applications, the event X is fixed in the discussion, and we wish to consider the impact of its having been observed on our belief in various possible events C_i. In such a situation the denominator of the last expression, the probability of the given evidence X, is fixed. For more on the application of Bayes' theorem under the Bayesian interpretation of probability, we can apply it in the application using Bayesian learning. we can apply the algorithms for the preference of item category based RFM using Bayesian theorem in previous paper.

For any set of random variables, the probability of any member of a joint distribution can be calculated from conditional probabilities using the chain rule (given a topological ordering of X) as follows:

$$P(X|C_i) = P(x_1, x_2, \ldots, x_n, C_i)\,P(C_i)$$
$$= P(x_1|C_i)\,P(x_2|C_i)\ldots P(x_n|C_i)\,P(C_i)$$
$$= P(C_i)\prod_{k=1}^{n} P(x_k|C_i) \qquad (3)$$

A belief network structure encodes the assertions of conditional independence in Equation 3, and it is a set of probability distributions corresponding to that structure. To compute prior/posterior probability, we typically have made the assumptions, which we explicate here. Assumption 1 The table of purchase is consisted of the sample data from the database of purchase data made by the behavior of purchase. Consider a domain U of User variables $\{u_1, u_2, u_3, u_4\}$, that is if a user(U_3) of site on e-commerce bought some brand items in item category of a domain C of brand item variables {C1, C2, C3, C4}, first, the system could suggest the brand item recommended by the preference of customer(a user(U_3) as prior probability using purchase counts {1, 1, 2, 2} and total purchase count has 6 counts. By the way, if a user(U_3) had 4 times of the behavior of purchase at the bran items (C4,), total counts is increased 10 purchase counts, then the system can provide the information of recommendation by the preference of customer(a user(U_3) as posterior probability of which is shown by table 1, the process of computation in Equation 4 after changing by the decision of the behavior of purchase[5].

$$P(C_3|X = C_4) = \frac{P(C_3)\,P(X = C_4|P(C_3))}{\sum_{i=1}^{4} P(C_i)\,P(X = C_4|C_i)} \qquad (4)$$

$$\frac{0.33 \times \frac{2}{10}}{0.17 \times \frac{1}{10} + 0.17 \times \frac{1}{10} + 0.33 \times \frac{2}{10} + 0.33 \times \frac{6}{10}} \cong 0.23$$

Table 1. Comparing Table of Preference

item u_3	C_1	C_2	C_3	C_4
Prior	$\frac{1}{6} = 0.17$	$\frac{1}{6} = 0.17$	$\frac{2}{6} = 0.33$	$\frac{2}{6} =$ **0.33**
Posterior	0.07	0.07	0.23	**0.65**

2.3. Clustering

Clustering is the process of organizing objects in a database into clusters. It involves classifying or segmenting the data into groups based on the natural structure of the data. Clustering techniques [6,7] 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. Clustering algorithm is a kind of customer's segmentation methods commonly used in data mining, can often use to k-means clustering algorithm. k-means is the most well-known and commonly, used partition methods are the simplest clustering algorithm. In the k-means algorithm, cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the cluster's center of gravity. 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 customer a between k measure in the kmeans algorithm[8]. The a_i means the preference of attribute i for customer a.

$$d_{a,k} = \sqrt{\sum_i (a_i - k_i)^2} \qquad (5)$$

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. In this paper, we can do clustering the purchase data to join customers' data using clustering of item category with weight, then we can use the groups of this cluster for the recommendation by customers' segmentation such as users' propensity, user's score and finally also forms the groups of customers with different features.

3. Proposing System

3.1. System Architecture

We can depict the system configuration concerning a personalized u-commerce recommender system using Bayesian learning and weighted preference under ubiquitous computing environment which is required by real time accessibility and agility. This system had four agent modules which have the analytical agent, the recommendation agent, the learning agent, the data mining agent in the internet shopping mall environment. We observed the web standard in the

web development, so developed the interface of internet to use full browsing in mobile device. As a matter of course, we can use web browser in wired internet to use our recommendation system. We can use the system under WAP in mobile web environment by using feature phone as well as using the internet browser such as safari browser of iPhone and Google chrome browser based on android so as to use our system by using smart phone.

3.2. Bayesian Learning and Weighted Preference

In this section, we can describe a personalized u-commerce recommender system using Bayesian learning and weighted preference. We can use the whole purchase data to join the customer information for pre-processing so as to be possible to recommend the items with efficiency. Then, the system can create the cluster with neighborhood user-group using the task of preprocessing for clustering purchase data to join customer's information with demographic variables: the code of classification such as age, gender, occupation, region, and FRAT variable. The system can take the preprocessing task using the whole purchase data, to apply the rate of weight based on the number of customers by the rank of customer's score and then create the cluster of purchase data sorted by item category, joined the cluster of user information called by customer DB, neighborhood user group. After that, the system can apply weight to the preference of brand item, that is, reflect the results of weighted preference by emphasizing the important item to recommend the item with high purchasability to be reflected frequently changing preference of customer as time goes by emphasizing the important item. The procedural algorithm of the Bayesian preference of item category is depicted as the following Table 2. The procedural steps using Bayesian learning and clustering of item category with weighted preference is depicted as the following Table3.

Table 2. The procedural algorithm of the Bayesian preference of item category

Step 1 : When the user joins the membership, user's information is created, managed the code of classification reflected demographic variable such as age, gender, occupation and skin type, as users' propensity.
Step 2 : The system can extract the purchase data with the login user's same propensity from the whole purchase data for pre-processing.
Step 3 : The system can set the purchase data in order by each user and by each brand item in item category.
Step 4 : The system can calculate the preference of item category as the rate of purchase.

$$P(C_j) = \frac{\sum_{i=1}^{n} f_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{k} f_{i,j}}$$

where i=1,2,3,...,n and j=1,2,3,...,k and $f_{i,j}$: number of purchase frequency

Step 5 : The system can calculate the preference of item category in order by each user in the cluster.

$$P(C_j \mid U_i) = \frac{P(C_j, U_i)}{P(U_i)}$$

where

$$P(C_j, U_i) = \frac{f_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{k}} \quad \text{and} \quad P(U_i) = \frac{\sum_{j=1}^{k} f_{i,j}}{\sum_{i=1}^{n}\sum_{j=1}^{k} f_{i,j}} \simeq P(C_{j,i})$$

That is, it's $C_{j,i}$ is calculated as the following expression.

$$P(C_{j,i}) = \frac{f_{i,j}}{\sum_{i=1}^{n} f_{i,j}}$$

Step 6: The login user's preference of item category as a prior probability is changed to a posterior probability through Bayesian learning, if an user wanted to buy any items as additional purchase.

$$P(C_{j,i} \mid X_i) = \frac{P(C_{j,i}, X_i)}{P(X_i)} = \frac{P(C_{j,i})\,P(X_i \mid C_{j,i})}{P(X_i)},$$

$$P(X_i) = \sum_{j=1}^{k} P(C_{j,i})\,P(X_i \mid C_{j,i})$$

where

Step7 : The system can recommend the brand item with the highest score in item category selected by descending order of the preference by each item category in the cluster with the login user's same propensity.

Table 3. The procedural steps using Bayesian learning and clustering of item category with weight

Step 1 : The FRAT score of customer is computed so as to reflect the attributes of the customer, consists of four attributes(F,R,A,T), each attribute has five bins divided by each 20%, exact quintile.
Step 2 : The system can count up all the number of customers by each rank of customer's FRAT scores, to make the table of the rate of weight.
Step 3 : The system can calculate the rate of weight based on the number of customers with each rank of FRAT score, then make the table of the rate of weight.
Step 4 : The system can scan whole database(sale) and calculate the weighted preference, weighted by each rank of FRAT score.
Step 5 : First, the system can recommend the item using prior probability according to the preference in the cluster selected by the code of classification reflected demographic variable and customer's FRAT score.
Step 6 : Then, the system could recommend the item according to the information of recommendation which is applied by posterior probability through Bayesian learning, if a customer wanted to buy any items as additional purchase

3.2. The Procedural Algorithm for Recommendation

The system can search the information in the cluster selected by using the code of classification reflected demographic variable and customer's FRAT score. It can scan the preference of brand item in the cluster, suggest the brand item in item category selected by the highest preference as the average of brand item. This system can show the list of recommendation with TOP-N of the highest preference of item as prior probability to recommend the item with purchasability efficiently, then the system could recommend the item according to the information of recommendation which is applied by posterior probability through Bayesian learning, if a customer wanted to buy any items as additional purchase. This system can recommend the items with efficiency, are used to generate recommending item according to the basic the weighted preference through clustering of item category with weight. It can recommend the associated item to TOP-N of recommending list if users want to have the cross-selling or up-selling. This system takes the cross comparison with purchase data in order to avoid the duplicated recommendation which it has ever taken.

4. The Environment of Implementation and Experiment & Evaluation

4.1. Experimental Environment

This system proposes a new recommending method using Bayesian learning and clustering of item category with weight based on FRAT under ubiquitous computing environment. In order to do that, we make the implementation for prototyping of the internet shopping mall which handles the cosmetics professionally and do the experiment. It is the environment of implementation and experiment below.

- OS: Windows
- Web Server: Apache HTTP Server Version 2.2.14 / WAP 2.0
- XML/WML2.0/HTML5.0/CSS3/JAVASCRIPT
- Server-Side Script: JSP/ PHP 5.2.12
- Database : MySQL Version 5.1.39
- Java 2 SDK(1.7.0_11)
- MySQL JDBC
- jQuery Mobile
- jakarta-tomcat (5.0.28)

We have carried out the implementation and the experiment for proposing system through system design, we have finished the system implementation about prototyping recommendation system. It could be improved and evaluated to new system through the result of experiment with MAE and the metrics such as precision, recall, F-measure as comparing proposing system using clustering of item category with weighted preference based on FRAT Score with other previous system using clustering of item category with weighted preference based on RFM score and existing system without weight.

4.2. Experimental Data for Evaluation

We used 319 users who have had the experience to buy items in e-shopping mall which handles the cosmetics professionally, 580 cosmetic items used in current industry, 1600 results of purchased data recommended in order to evaluate the proposal system. It could be evaluated in MAE and Precision, Recall, F-measure for the recommendation system in clusters. It could be proved by the experiment through the experiment with learning data set for 12 months, testing data set for 3 months in a cosmetic cyber shopping mall. We'd try to carry out the experiments in the same condition with dataset collected in a cosmetic internet

4.3. Experiment & Evaluation

The proposing system's overall performance evaluation was performed by dividing the two directions. The first evaluation is mean absolute error(MAE). The mean absolute error between the predicted ratings and the actual ratings of users within the test set. The mean absolute error is computed the following expression (6) over all data sets generated on purchased data.

$$MAE = \frac{\sum_{i=1}^{N} |\epsilon_i|}{N} \quad (6)$$

N represents the total number of predictions, ε represents the error of the forecast and actual phase i represents each prediction.

Table 4. The result for table of MAE by comparing proposal system with existing system

	P_count	Proposal	previous	Existing
MAE	50	0.34	0.47	0.65
	100	0.19	0.23	0.32
	300	0.06	0.07	0.08
	500	0.04	0.05	0.06

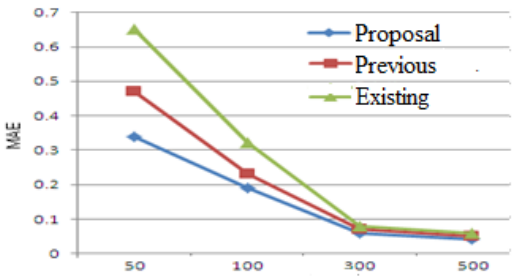


Figure 1. The result for the graph of MAE by comparing proposal system with existing system

The next evaluation is precision, recall and F-measure for proposing system in clusters. The performance was performed to prove the validity of recommendation and the system's overall performance evaluation. The metrics of evaluation for recommendation system in our system was used in the field of information retrieval commonly [10].

Table 5. The Result for Table of Precision, REcall, F-Measure for Recommendation Ratio by Each Cluster

Cluster No.	Proposal(W_FRAT)			Previous(W_RFM)			Existing		
	Precision1	Recall 1	F-measure1	Precision2	Recall 2	F-measure2	Precision3	Recall 3	F-measure3
C1	47.12	81.22	59.64	46.82	80.66	59.25	56.98	50.89	50.21
C2	47.36	64.73	54.70	48.00	55.27	51.38	38.97	15.18	20.88
C3	54.52	82.86	65.77	52.67	79.26	63.28	42.08	16.07	22.34
C4	43.50	81.95	56.84	42.23	80.66	55.44	48.79	31.32	35.64
C5	63.30	94.64	75.86	55.91	92.59	69.72	50.60	13.88	21.03
C8	42.92	65.23	51.77	42.81	64.52	51.47	52.49	34.98	39.75
C7	48.57	78.57	60.03	41.58	74.08	53.26	47.41	26.81	32.26
C8	31.09	78.57	44.55	27.72	74.08	40.34	46.68	25.19	30.28
C9	49.94	86.47	63.32	46.39	85.19	60.07	46.53	18.32	25.10

The p-count is the number of count for purchase data based on the number of customers

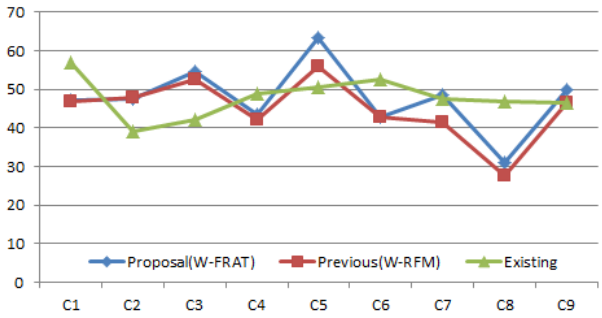


Figure 2. The result of recommending ratio for recommendation each cluster by precision

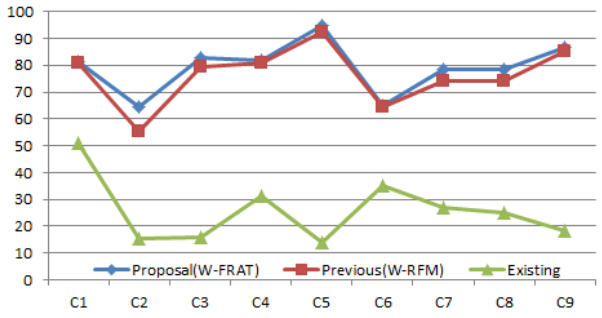


Figure 3. The result of recommending ratio for recommendation each cluster by recall

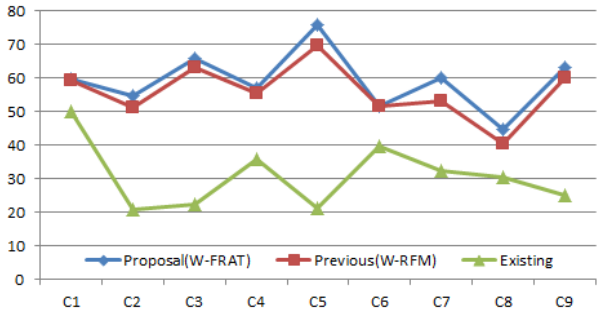


Figure 4. The result of recommending ratio for recommendation each cluster by F-measure

Above table 5 shows the result of evaluation metrics (precision, recall and F-measure) for recommendation system. It shows the improvement in the result of evaluation rates for proposing system comparing with both previous system and existing system. The proposed is higher 30.31% in precision, 13.41% in recall, 17.76% in F-measure than the existing system. As a result of that, the performance of the proposal system is improved better than both previous system and existing system. The following figure 5 shows cosmetic items on the web of a recommendation system using Bayesian learning and weighted preference and also, a smart phone is available to show that. This system can be used immediately in u-commerce under ubiquitous computing environment which is required by real time accessibility and agility after finishing a particular tasks such as Bayesian learning and weighted preference through the task of calculating weighted preference and clustering of item category for pre-processing to reduce the processing time.

Figure 5. The result of recommending items of cosmetics

5. Conclusion

Recently u-commerce as an application field under ubiquitous computing environment required by real time accessibility and agility, is in the limelight[4]. We proposed a personalized u-commerce recommender system using Bayesian learning and weighted preference in order to reflect frequently changing preference of customer as time goes by emphasizing the important item, to improve the accuracy of recommendation with high purchasability. The existing system did not reflect the importance of the item, and do not consider these dynamic changes of the preference of customer as time goes by emphasizing the important items in item category. It is crucial to have different weights for many different items and reflect the results of weighted preference. We have described that the performance of the proposing system using Bayesian learning and weighted preference is improved better than both the previous system (W-RFM) and existing system. To verify improved better performance of proposing system, we carried out the experiments in the same dataset collected in a cosmetic internet shopping mall. It is meaningful to present a personalized u-commerce recommender system using Bayesian learning and weighted preference under ubiquitous computing environment. The following research will be looking for ways of recommending service using an implicit method without onerous question and answer under ubiquitous computing environment.

Acknowledgment

This research1) 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. References

[1] Cho, Y.S., Jeong, S.P., Ryu, K.H.:" Implementation of Personalized u-commerce Recommendation System using Preference of Item Category based on RFM," *In: The 6th International Conference on Ubiquitous Information Technologies & Applications*, 2011, pp. 109–114

[2] Cho, Y.S., Moon, S.C., Noh, S.C., Ryu, K.H.: "Implementation of Personalized recommendation System using k-means Clustering of Item Category based on RFM," *In: 2012 IEEE International Conference on Management of Innovation & Technology*, 2012

[3] Cho Y.S., Moon S.C., Jeong S.P., Oh, I.B., Ryu K.H., "Clustering Method using Item Preference based on RFM for Recommendation System in u-Commerce," *Ubiquitous Information Technologies and Applications Lecture Notes in Electrical Engineering, Springer*, Volume 214, 2012, pp. 353-362.

[4] Cho, Y.S., Moon, S.C., Jeong, S.P., Oh, I.B., Ryu, K.H.: "Clustering Method Using Weighted Preference Based on RFM Score for Personalized Recommendation System in u-Commerce," *In: Y.-S. Jeong et al. (eds.) Ubiquitous Information Technologies and Applications. LNEE, Springer, Heidelberg*, 2014, vol. 280, pp. 131-140.

[5] Choi., J. H., Kim., D. S., Rim., K. W., "Dynamic Recommendation System for a Web Library by Using Cluster Analysis and Bayesian Learning," *KCI*, Vol. 12, No 5, 2002, pp. 385-392.

[6] Hand, D., Mannila, H., Smyth, P.: "Principles of Data Mining," *The MIT Press ,2001*

[7] Collier, K., Carey, B., Grusy, E., Marjaniemi, C., Sautter, D.: "A Perspective on Data Mining," *Northern Arizona University*,1998

[8] Shin, M.-S.: "An Alert Data Mining Framework for Intrusion Detection System", *Journal of Korea Academia-Industrial cooperation Society*, 12(1), 2011

[9] Hastie, T., Tibshirani, R., Friedman, J.: "The Elements of Statistical Learning – Data Mining, Inference, and Prediction", *In: Springer*, 2001

[10] Herlocker, J.L., Kosran, J.A., Borchers, A., Riedl, J.: "An Algorithm Framework for Performing Collaborative Filtering," *In: Proceedings of the 1999 Conference on Research and Development in Information Research and Development in Information Retrival*, 1999

Distance Metric Learning for Face Recognition

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

Abstract

Most machine learning algorithms strongly depend on a distance metric. For example, an appropriate distance metric is crucial for k-Nearest Neighbor (kNN) classifier that does not require a training phase. Learning the metric from a given training set can improve the performance of the machine learning algorithms instead of just using Euclidean distance. Recently, distance metric learning methods that is trained with a given data set have reported significant improvement of the kNN classifier. In general, the performance of a distance metric learning method is variable for each application. The DML methodsZA can be categorized into two groups; supervised and unsupervised methods. The supervised method learns the distance metric using the training data with their class label. The unsupervised method learns the distance metric using the training data only. Specifically, the goal of the supervised method is to compute the covariance matrix of a Mahalanobis distance metric.

In case of the unsupervised method, it finds out a linear transformation. In this thesis, we review both methods and focus on the supervised methods for face recognition. Then, we compare the performance of the state-of-the-art distance metric learning methods; Principal Component Analysis based methods, Neighbor Component Analysis, Large Margin Nearest Neighbor, and Energy-based method. To find out a proper method for face recognition, we did several experiments on the public face database; ORL, PIE, and YaleB. Each method achieves its best performance with different parameters from those of others. Therefore, we evaluated all the number of dimensions and parameters to find out the proper parameters for each method and database. For overall performance comparison on three databases, we analyzed both the correlations and t-test on the error rates of all the methods and databases. We also evaluated the time complexity of them with the same dimension. 00Our experimental results on the public face databases demonstrate that the Mahalanobis distance metric based on PCA is still competitive with respect to both performance and time complexity in face recognition.

Design for Triple Helix Model Framework using Information of Bibliography

Gwi Suk Gim, Ho Sun Shon, Byung Jun Cho, Hyung Chul Rah, So Young Kim
Graduate School of Health Science Business Convergence,Chungbuk National University
{ggsuk.psm, shon0621, cho135135, rah.remnant, letter.sykim}@gmail.com,

Abstract

Globally, during the intensified competition of science and technology, the goal of the country is to create economic benefits through the development and utilization of new technologies. In this knowledge-based society, linkage of University-Industry-Government is important for technological advances. Linkage structures of University-Industry-Government usually were analyzed by Triple Helix model. Triple Helix model describes the structure of the knowledge-based innovation system in knowledge economics. Using this model, we can analyze development capability and predict research trends. In this paper, we designed Triple Helix model framework using bibliographical information analysis.

1. Introduction

Today, in the international community, each country's goal is the creation of economic interests through development and utilization of new technologies as a promising area. However, now as in a knowledge-based society, it is too hard to develop high technology without linkages and synergies of University-Industry-Government.

Knowledge infrastructures in knowledge-based innovation systems were described in terms of the network approach. The network approach is a method to identify the structure of the society system based on the relationship among the elements of the system rather than the properties of the each object[1].

Triple Helix model is commonly used to describe the structure of the knowledge-based innovation system in knowledge economics[2]. Triple Helix model was used to analyze and track the dynamics of the research system of some countries.

We can predict the future trends through analysis of national research system and identify technology capability of all the countries of the world.

When data was analyzed by triple helix method, bibliographic information data analyzed by triple helix-based method is mainly used in bibliometrics. It required Triple Helix model which provides to analyze and to standardize a number of bibliographical information.

So, in this paper, we designed Triple Helix model framework using bibliographical information analysis.

2. Related Work

The term of Triple Helix was formed in order to emphasize that universities, industry, and government influence each other closely while receiving the triple spiral shape to participate in knowledge production and innovation like a double helix form of DNA.

Triple Helix thesis was formulated a scholastic exchange between Etzkowitz and Leydesdorff in 1994[3]. Etzkowitz is professor in U.K. who was steadily interested in research of university-business relationship. And Leydesdorff is professor in Dutch who focused on the evolutionary models to generate hyper cycle.

Alan poter in Georgia Institute of Technology in USA was built Vantage Point software which using text mining concept[4]. It analyzes bibliographic and R&D database quickly, then creates charts and graphs to extract relevant information for help better decisions.

Leydesdorff made a calculation program for Triple Helix[5].

3. Triple Helix Model

Triple Helix indicator which is based on the mathematical theory of communication methods used analysis for University-business-government cooperative system to organic. It uses methodology developed by Leydesdorff in order to measure the degree of dynamism in the network between University-Industry-Government[6].

In Triple Helix model, measuring T value of cross-linking is derived from "A mathematical theory of communication" based on the second law of thermodynamics, the entropy of physics concepts [7].

The entropy calculation formula is as follows:

$$H=-\sum_{i=}^{n} P_i \log_2 P_i \quad (1)$$

The following expression is case of university

$$H_u=-\sum_{u} P_u \log_2 P_u \quad (2)$$

Here, P is the probability, and u means college. Also i is the enterprise, g is the government, and H represents the entropy.

T (transmission) is the amount entropy which is the probability of increasing university-business-government interaction, and the transfer of information in the network level, in other words, is Triple Helix indicators

Therefore, when T value is negative, the entropy decreases. However, the dynamics of the network between university-business-government increases.

$$Tuig = Hu + Hi + Hg - Hui - Hig - Hug + Huig \quad (3)$$

All values are represented by bit which is the basic unit of information.

Triple Helix data analysis is mainly used in bibliometrics based on information from the paper.

Therefore, it uses bibliographical information to analyze the activity of mutual knowledge communication with university-business-government.

UI = the number of joint paper with university and industry.

UG = the number of joint paper with university and government research institutions.

IG = the number of joint paper with industry and government research institutions.

UIG = the number of joint paper with university,, industry and government research institutions

U, I, G = the number of own thesis each universities, industry, government research institutions

UI, UG, IG, UIG, U, I, and G are independent values.

4. Triple Helix Model Framework

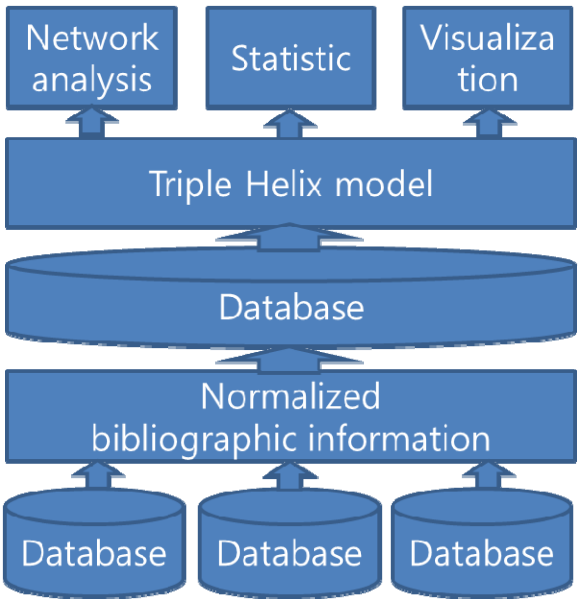


Figure 1. Design of Triple Helix model framework

This figure 1 is triple helix model framework designed in this paper.

First, we get bibliographic information from databases as a Web of Science, SCOPUS, and Google Scholar. Then this unstructured data is normalized based on organization address. It uses pattern matching of text mining, for example. When comparing the address of data, we converted it to the full name of organization from the name of organization as abbreviation if they are matched. And we build integrated database using normalized data. Triple Helix model is standardized bibliographic information classified as University-Industry-Government and calculated each activity values of mutual knowledge communication, i.e., entropy. Finally, we provided network analysis, statistic and visualization of results.

The proposed Triple Helix model framework normalized unrefined bibliographical information in conventional methods before the data is analyzed. It can provide more accurate and a user-centered convenient analysis because it contained the data cleaning process in this program.

5. Conclusion and Further Work

In this study, we designed Triple Helix model framework which uses bibliographical information in order to analyze the technical development capability of each country. It standardized unstructured bibliographical information as organization name.

Then we analyze the interaction among university, industry, and government. Thus it will give high performance such as processing speed and accuracy of analysis.

Also, we will implement Triple Helix model framework in future research. When we make systematization of this framework, it will be applied to help the National Science and Technology Policy decisions using analyzed development capability of the country and predicted future trends.

Acknowledgment

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), ISBB (International *Science* Business Belt**)** support program (2013K001552), and Basic Science Research Program through the National Research Foundation of Korea (N RF) funded by the Ministry of Science, ICT & Future Planning (No-2013R1A1A206518).

6. References

[1] S. Wasserman, G. Joseph, eds., “Advances in social network analysis: Research in the social and behavioral sciences.”, *Sage Publications*, Vol. 171., 1994.

[2] L. Leydesdorff, H. Etzkowitz, "The triple helix as a model for innovation studies.", *Science and public policy*, 25.3 ,1998, pp.195-203.

[3] H. Etzkowitz, "Academic-industry relations: A sociological paradigm for economic development.", *Evolutionary Economics and Chaos Theory: New directions in technology studies*, 1994, pp.139-151.

[4] A. L. Porter, J. Michael, Detampel, "Technology opportunities analysis.", *Technological Forecasting and Social Change*, 49.3, 1995, pp.237-255.

[5] http://www.leydesdorff.net/th/th.exe

[6] L. Leydesdorff, "The mutual information of university-industry-government relations: An indicator of the Triple Helix dynamics.", Scientometrics, 58.2, 2003, pp.445-467.

[7] C. E. Shannon, "A mathematical theory of communication.", ACM SIGMOBILE Mobile Computing and Communications Review, 5.1, 2001, pp.3-55.

Framework Design for Big Data Analysis in Patients with Acute Myocardial Infarction

Changwoo Woo, Wooyeong Jang, Ho Sun Shon, Eung-Do Kim, Gilwon Kang
Graduate School of Health Science Business Convergence, Chungbuk National University, South Korea
{cwwoo.psm, jjangwy8838, shon0621, trlfighting, gilwon67} @gmail.com

Abstract

Heart disease is a leading cause of death in four of Korean. Also, heart disease is currently No.1 cause of death in the U.S. Korea has increased the incidence of Acute myocardial infarction which is the leading cause of cardiovascular disorders. Myocardial infarction (MI) or acute myocardial infarction (AMI) is the medical term for an event commonly known as a heart attack. MI occurs when blood stops flowing properly to a part of the heart, and the heart muscle is injured because it can't get enough oxygen. Usually this is because one of the coronary arteries that supply blood to the heart blocks the flow of blood due to an unstable buildup of white blood cells, cholesterol and fat.

If this problem can't be solved early, patients with acute myocardial infarction death tend to have a higher fatality rate. Acute myocardial infarction also leads to serious complications if it's not treated. So it is very important to diagnose and treat the patients as early as possible. Because there include a lot of structured and unstructured data of patients with acute myocardial infarction, database applied to big data platform skill is required.

In this paper, we designed frameworks based big data environment for various data analysis about myocardial infarction patient. We contrived predictive models of patients with acute myocardial infarction by developing each modules about designed framework. We expect risk of recurrence of patients discharged after treatment.

1. Introduction

In Korea, cardiovascular disease caused by atherosclerosis is steadily increasing, because of changes in diet and lifestyle habits. According to census data about cause of death in National Statistical Office, cause of death from cardiovascular disease and cancer has been reported as the leading cause as 47.1% of all [1].

Main disease of arteriosclerotic heart disease is myocardial infarction and angina. MI occurs when there is a diminished blood supply to the heart which leads to myocardial cell damage and ischemia [2, 3].

The prevention of the disease is important for patients with acute myocardial infarction. For the treatment of the disease, first aid and any other necessary medical treatment is very important. So, we need analysis model to analyze the data.

But structured data and unstructured data include various properties. So, another big data need to differ from original data base.

In this paper, we designed frameworks based various data analysis about acute myocardial infarction patient.

2. Related work

Recently, IBM Big Data was used to predict heart disease. IBM, Sutter Health, Geisinger Health Systems found the signs of heart disease through big data analysis [4].

Subject of the experiment is analysis of health records such as taking the drug, medical history, and genetic analysis. So, we need to search the common factors that cause heart disease.

Recently, big data processing and analysis of medical data are important and emphasizing issues. For example, in the case of heart disease, big data analysis can be applicable to judge whether or not stent technology is applicable [5].

In IBM and medical insurance company's WellPoint, doctors and other medical staff offer diagnosis and treatment to the patient through big data analysis. Integrated data and information of health insurance

registered in the 34M people search complex medical treatment [6].

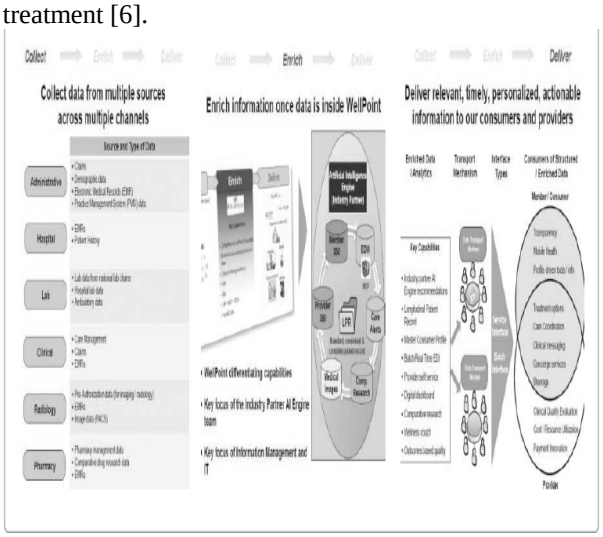


Figure 1. Data collection, enhancement, delivery process of WellPoint.

3. Framework for big data process

We contrived and designed the framework about big data analysis of acute myocardial infarction patients.

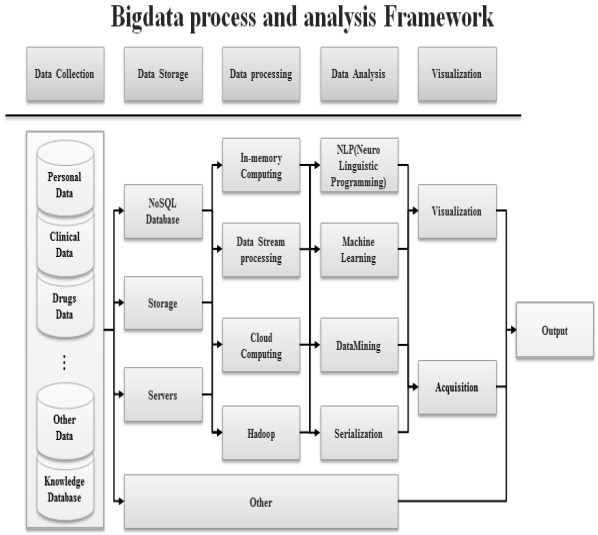


Figure 2. Analysis of patients with acute myocardial infarction based framework for Big data.

Data collection scattered in various locations is unified into one database about personal, clinical, drug.

In data storage step, the consolidated data is stored into NoSQL DB or the Storage or Server.

In data processing stage, data is processed by In-Memory computing and data stream method in real-time or hadoop and cloud computing in a distributed data.

In data analysis stage, data patterns were founded by natural language processing, machine learning method or analysis using data mining and serialization.

Finally, in visualization stage it is processed by visual expression as chart or graphic about data analysis in order to easily pass data to user. Acquisition stage is processes for reinterpretation of the data.

In this paper, if you save a typical DBMS with various properties of data, the business model designed by the data framework can be uniformly reduced to consumption of resources about patients with acute myocardial infarction.

Table 1. The difference between old database and database based on Big data.

	Legacy DBMS	Bigdata DBMS
processing speed(velocity)	low	high
Data Size(volume)	small	large
Unstructured Data(variety)	Not supported	support

4. Conclusion

This study deviates from DBMS of the traditional method for the various analyses of patients with acute myocardial infarction. Also, we designed frameworks for analysis of big data environment. big data-based database can resolved difficult to deal with unstructured data including basic existing DBMS. It will be able to show high performance in various sectors such as processing speed and data size and so on.

In the future, we will contrive the predictive models of patients with acute myocardial infarction by developing each modules about designed framework. We expect risk of recurrence of patients discharged after treatment.

This processing system of big data can be established by close connection among the patient, hospitals and insurance companies.

Therefore, if IT technologies are introduced in order to reduce health care costs, online health care and various online web services would be utilized.

Acknowledgment

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), ISBB **(**International *Science* Business Belt**)** support program (2013K001552), and Basic Science Research Program through the National Research Foundation of Korea (N RF) funded by the Ministry of Science, ICT & Future Planning (No-2013R1A1A206518).

5. References

[1] National Statistical Office, “Cause of death statistics”, 2012

[2] http://www.komir.org

[3] Ki hong lee, et al, “Sex Differences of the Clinical Characteristics and Early Management in the Korea Acute Myocardial Infarction Registry”, *Korean Circulation J*, 37, 2007, pp. 64-71

[4] http://venturebeat.com/

[5] http://www.nims.re.kr/

[6] National Information Society Agency, "10 global best practices big data - big data lead to the world", 2012, pp.26-29

[7] Chul Kim, “Cardiovascular diseases and sports medicine”, *Korean Med Assoc*, 54(7), 2011, pp.674-684

Session : PSM, Text Mining & Natural Language Processing

- Invited speakers
  Wang Ling
- Keyword Extraction using Anti-pattern
  Khuyagbaatar Batsuren, Tsendsuren Munkhdalai, Meijing Li, YoungJung Kim, Jong Yun Lee
- Developing Graph Database for Multilingual Corpus
  Hyeon Ah Park, Khuyagbaatar Batsuren, Nak Hyun Choi, Jeong Hee Hwang
- Classification of Diseases from Number of Outbreaks
  Wooyeong Jang, Changwoo Woo, Ho Sun Shon, Young-Sung Lee, YoungGyu Kim, Keun Ho Ryu
- Development of Web-based System for Analysis of Urinary Cancer Patient from Disease Prevention Questionnaire
  Kyeong Seok Lee, Hyun Woo Park, Soo Ho Park, Kyung Ah Kim

Keyword Extraction Using Anti-pattern

Khuyagbaatar Batsuren[1], Tsendsuren Munkhdalai[1], Meijing Li[1], Youngjung Kim[2], JonYun Lee[2]
[1]*Database/Bioinformatics Laboratory, Department of Computer Science,*
Chungbuk National University, South Korea
{huygaa, tsendeemts, mjlee, khryu}@dblab.chungbuk.ac.kr
[2]*Department of Digital Informatics and Convergence, Chungbuk National University, Korea*
{rex,jongyun}@chungbuk.ac.kr

Abstract

Keyword extraction has been utilized to improve many applications such as document summarization and clustering. Despite of a number of studies for graph-based keyphrase extraction, all of them use only the word co-occurrence relation graph. This paper introduce a novel graph named dependency graph and anti-pattern. Motivated by them, we proposed a novel graph-based keyword extraction using anti-pattern. An anti-pattern which usually appears in non-keyphrases is used to filter out candidate phrases. For experimental works, we have conducted the experiments showing comparison between a co-occurrence graph and dependency graph, compare with other existing methods for keyphrase extractions, and find out the experimental results of how anti-patterns impact and improve the performances of unsupervised keyphrase extraction.

1. Introduction

Text document understanding, summarizing, and information retrieving required that this kind of techniques must always be better than before because of size of text information have been increasing dramatically. A keywords can be considered as brief summary of a text document. Consequently, a machine can easily understand them for some task. Therefore, keyword extraction have been utilized to improve many applications of Natural Language Processing and Text mining. For example: a document clustering and summarization information retrieval. The task of keyphrase extraction is to automatically identify in a text a set of terms that best describe the document.

Even though keyphrases of research articles are usually assigned by authors, many other documents such as the news articles, review articles don't have author assigned keyphrases. Despite a number of previous studies for keyphrase extraction, performances of them aren't perfect as other area such as the pos tagger, named-entity recognition. So the task of keyphrase extraction is still an important topic. Previous works can be generally categorized into either unsupervised and supervised methods. Supervised methods first extract a features from candidate phrases for training and supervised machine learning algorithms have been used to learn the model on extracted features. Finally, the model classifies a candidate noun phrase either keyphrase or not. In [9-12] works of supervised learning, Naive Bayes, Decision Tree, and Neural Network machine learning techniques have been widely used. Weakness of supervised method is that requires a big human-annotated datasets and the trained model can only work well on the trained area. Since TextRank, which is graph-based ranking algorithm using word co-occurrence relation, was proposed by Mihalcea, most of unsupervised methods exploit and extend his invaluable idea and developed many variants of Textrank. These approaches are state-of-art methods for keyword extraction.

Most approaches for keyword extraction have a one common stage, which is a generation of candidate keyphrases. Supervised methods, based on Naive Bayes and Neural Networks, and unsupervised methods rank candidate keyphrases and top ones are selected as the keywords. Moreover, although a many number of studies for keyword extraction were proposed, our proposed new ideas aren't used in any other work. We introduce a novel graph, called the dependency graph, and an anti-pattern. Our research is based on the combination of anti-pattern and graph-based unsupervised ranking algorithm Textrank. For our study, a dependency graph are used instead of using a word co-occurrence relation-based graph on Textrank. When comparing the dependency graph with a state-of-art graph, this graph aren't caught in any window size and

based on grammatical relations. Furthermore, some opportunities have been opened by using the dependency graph and many sentence structural information can be obtained from the graph. These opportunities can be very useful for the definite task such as keyword extraction.

After generating candidate noun phrases, most of them are non-keyword. After long-term monitoring them, many non-keyphrases are known with same patterns. Therefore, we extracted these patterns called anti-pattern which often occurs in non-keyphrases. When using anti-patterns to prune candidate phrases, performances of our method increased unbelievably. The our purposed method combined advantages of both supervised and unsupervised methods to improve a state-of-art keyword extraction. Hence, anti-patterns contain some general advantages of supervised method.

The rest of this paper is organized as follows: In section 2 we will review some related works. Section 3 contains more details about our proposed a novel ideas and method: Dependency graph and Anti-patterns. Subsequently, A Novel Graph-based Keyphrase Extraction using Anti-pattern is explained. In section 4, the experimental analysis, discussion and the related results are provided. Last section addresses our conclusion.

2. Related works

The proposed method is based on the combination of anti-pattern and graph-based unsupervised ranking algorithm Textrank [6]. When learning the anti-patterns, it requires a training data set. Therefore, our work is actually related to both unsupervised and supervised keyphrase extraction. Graph-based ranking methods are state-of-art in unsupervised keyphrase extraction. Mihalcea et al [6] first applied a graph-based ranking algorithm (TextRank) for keyword extraction. TextRank was inspired by Pagerank by using the ranking algorithm for a text and builds a graph representing a text. Every node V_i corresponds to a lexical unit. The goal is to calculate the score of each node $WS(V_i)$ which reflects its importance, and then adopt the words types that correspond to the highest-scored vertices to form keyphrases for a given text. $WS(V_i)$ is initialized with a default value and computed in an iterative manner as follows a recursive formula.

$$WS(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} \frac{w_{ij}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j) \quad (1)$$

where w_{ji} is the weight of direct edge (V_j,V_i), $In(V_i)$ is the set of vertices that point to vertex V_i , and $Out(V_j)$ is the set of vertices that vertex V_j points. d is the damping factor usually set to 0.85, as in the PageRank algorithm.

Many unsupervised methods for keyword extraction are recently proposed. Most of them exploits the idea of Textrank. ExpandRank [7] is a Textrank extension which uses a small number of nearest neighbor documents to provide more knowledge to improve keyphrase extraction. After finding k nearest neighbors documents of the document using Tfidf and cosine similarity, the graph for the document is built using the their similarities and co-occurrence statistics. Once the graph is constructed, the rest of the procedure can be similarly performed as Textrank. A topic decomposition framework was proposed by Liu et al [8]. Its first recognized a topic distribution from the dataset using Latent Dirichlet Allocation (LDA) model in [9]. Extracted topic distribution has a number of topics, each of which related to a group of words. In their work, multiple random walks are performed for the topics instead of the traditional single random walk through the graph. After this work was published, topical keyphrase extraction methods [10-11], which are quite similar to it, were applied to Twitter dataset.

A number of supervised machine learning method in this area has been proposed to classify a candidate phrase into either a keyphrase or not. Keyphrase extraction program called KEA was developed by Frank et al. [4-5], uses a Bayesian classifier. A Neural Network based approach in [2-3] has been presented. Medelyan and Witten propose KEA++ that enhances automatic keyphrase extraction by using semantic information on terms and phrases gleaned from a domain specific thesaurus. Before training a model of all the above supervised methods, the features are extracted from the training data set. Moreover, the most important features are the frequency and location of the phrase in the document. More linguistic knowledge has been explored by Hulth et al. [12]. Nguyen and Kan presented keyphrase extraction in scientific articles by using features that capture the logical position and additional morphological characteristics of scientific keywords. Keyword extraction using Naive Bayes in the medical domain has been presented in [13], which has been tested on a small set of 25 documents. These studies have been utilized in many different domains such as medical domain, computer science articles, web pages, and news. When comparing unsupervised methods for keyword extraction with supervised methods, the performances of unsupervised methods are relatively worse than a supervised methods. However, a

supervised method with high performances requires a big training corpus with a keywords assigned by a human. But a size of the training corpus with a keywords is actually very rare. Furthermore, these corpuses are only covering a specific domain. Therefore, a model for a specific domain that it is trained on, can only recognize features of keyphrases for this domain. Moreover, if it is used on another domain, it reaches a poor result. Therefore, in this paper, we propose a novel graph-based keyphrase extraction method using dependency parsing and anti-patterns.

3. Dependency graph and Anti-pattern

In this paper, we introduce the dependency graph and the anti-patterns that our proposed method is based on. Let us first define what the dependency graph is and how it can be constructed, and then explain what the anti-patterns are, categorize the different types of the anti-patterns, and how it can be measured.

3.1 Dependency graph

In order to apply graph-based ranking algorithms on a text, we need to construct the graph for a text. The Dependency Graph represents the text document and interconnects words with grammatical relation. The text is split into a tokens, which represent the nodes of the graph. After parsing all typed dependency relations from all sentences, a nodes are connected with weighted and directed edges based on typed dependency relations. A value of the edge is a sum of values of dependency relations between its nodes. A process of building dependency graph consists of following main steps:

- Parse a typed dependency relations for each sentence
- Identify terms that best define the proposed task and add them as vertices in the graph
- Draw edges between vertices in the graph using these relations. Edges are directed and weighted.

Stanford Dependency parser is utilized to extract grammatical relations between words in a text. It provides a representation of grammatical relations between words in sentence. The current representation contains approximately 50 grammatical relations. Moreover, Stanford dependency parser has four kinds of models. In our research approximately 30 collapsed dependencies are utilized because the collapsed dependency summarize two basic dependencies and some dependencies aren't necessary for our task. Therefore, these unnecessary dependencies such as det, predet, aux, and advmod aren't used. In figure 2, the sample dependency graph for an abstract from Inspec dataset is shown. The sample abstract is "Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types." that same as a example abstract on Mihalcea's study [1].

The basic dependency graph can be modified for the task of keyword extraction. In order to improve the performances of our method, we modified the basic dependency graph. The vertices of the graph are categorized into noun, adverb, adjective, adverb, and

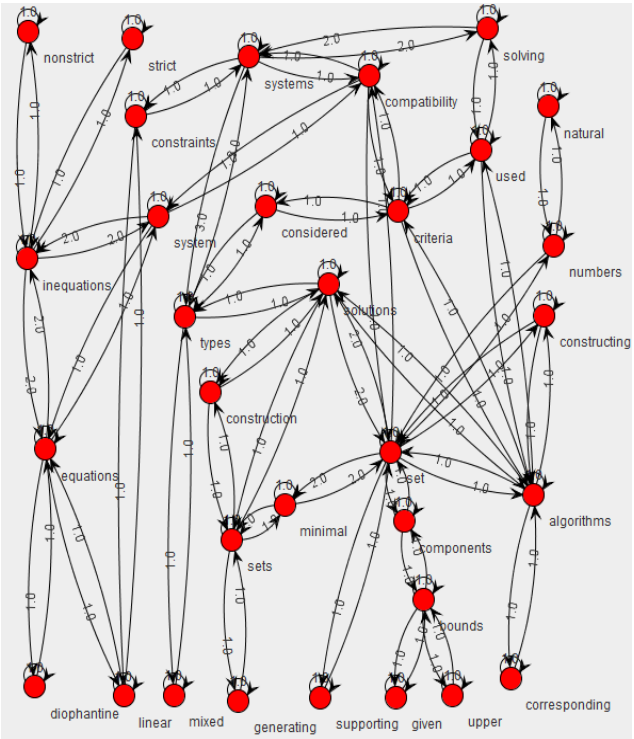


Figure 1. Basic Dependency Graph

verb vertices. Moreover, nouns and adjectives are more important than determiner, adverbs and verbs due to keyphrases are consists of adjectives and nouns. Therefore, the purpose of our modification is to consider the noun and adjective nodes as more important than other types of nodes. This modification are consists of following steps:

1. Assign the unique id for every verb and adverb in each sentences before constructing graph, and

assume each verb and adverb in each sentences are different than each other although verbs with same value and meaning are used in two sentences. Finally, add them as vertices in the graph.

2. Set the value of term frequency to the value of directed edge which connect a node to itself.
3. Draw edges between vertices in the graph using these relations. Edges are directed and weighted.

After processing first step, the importance values of the adverb and verb nodes in the graph would be decreased and it helps to increase an importance of the noun nodes. At second step, the noun and adjective nodes have their term frequency (TF>=1). However, all adverb and verb nodes have the term frequencies of 1 due to every adverb and verb in the document are different than each others. By doing step 3, the noun vertices only recommend other noun and adjective vertices as an important, and the adjective vertices only recommend noun vertices because of adjective words can only connect noun words in dependency tree. After this modification, the importance values of the noun and adjective nodes would be increased and the importance values of other types would be decreased. In Figure 2, the example of the modified dependency graph is shown.

3.2 Anti-pattern

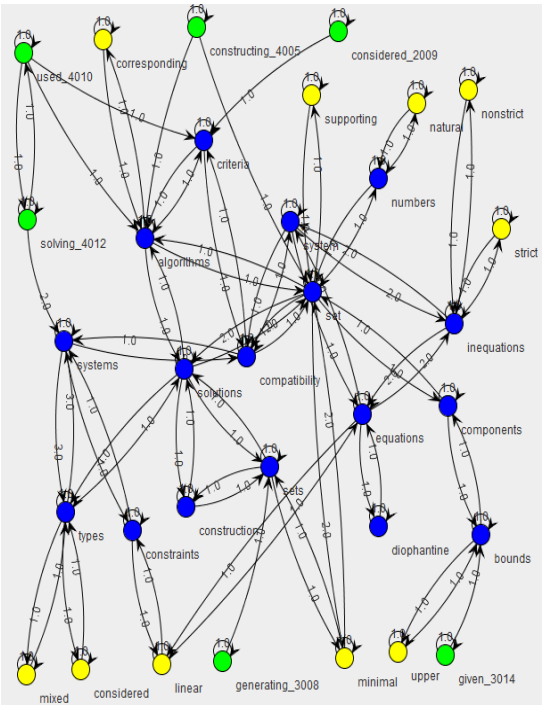


Figure 2. Modified Dependency graph

Anti-pattern is pattern that are usually occurred in non-keyphrases and are used to recognize a definite non-keywords from candidate keyword list. By detecting a definite non-keywords from a candidate phrases, the performances of keyphrase extraction methods will be increased. Anti-pattern can be classified into four types: head word, tail word, single word, and anti-word.

Table 1. Anti-pattern

	Type	Description	*c*
1	Head word	first word of candidate phrase	5
2	Tail word	last word of candidate phrase	5
3	Single word	candidate phrase with one length	2
4	Anti-word	any word of candidate phrase	8

The four kinds of anti-patterns are illustrated in Table 1. Let's consider this example: "new ranking algorithm" and "new method" phrases are generated in a candidate list and both classes of them are "No" then we can learn a "new" head word for anti-pattern. Then, a candidate phrases, a first word of which is new, are pruned from the candidate phrase list by using a "new" head word for anti-pattern.

Many single words are recommended as keyphrases after generating candidate noun phrases. Most of them aren't usually keywords. So Learning process provides many pattern with one length such as "system", "method". The four kinds of anti-patterns are trained as described in Table 1.

When measuring a strengths of anti-patterns, we used following formula:

$$Power(X) = \frac{P(Class = "Keyword" | X)}{P(X)}, (|X| > c) \quad (2)$$

X is pattern, |X| is representing number of candidate noun phrases which matches X. And c is threshold. If Power(X) is near by a zero, it is strong pattern. When learning anti-pattern, it requires training set. Although anti-patterns are only covering a training domain, some strongest pattern can help to prune non-keyphrases of other domains.

4. Proposed methods

The goal of our study is to improve state-of-art approach for keyword extraction. Our proposed method consists of three primary components: a candidate phrase filtering, a term weighting, and a post processing. At first stage of the candidate phrase filtering, all candidate noun phrases are generated from the dataset.

Then Anti-patterns are trained from candidate phrases for the training in order to filter a candidate keyphrases for testing dataset. For term weighting, the dependency graph is first built using dependency parsing. Once the graph is constructed, the random walk algorithm as TextRank are iterated until a convergence. For post processing, top-ranked phrases are selected as keywords. On the Figure 3, Flowchart of our method is displayed.

4.1 Candidate phrase filtering using Anti-patterns

For preprocessing phase, LingPipe sentence extractor are first used to detect sentences from a text document. Then Stanford dependency parser are utilized to extract the relations from each sentences separately in document. Even though Stanford dependency parser has the library to tokenize sentences on a text, it is unsupervised traditional approach. Moreover, LingPipe sentence extractor is supervised method based on the large training corpus. Before extracting relations, sentence structure and proper grammar of a document are very important. Therefore some preprocessing techniques such as tokenization, stemming, removing stop words mustn't be used on a document. After extracting a dependency relations, the dependency graphs are constructed as described in the dependency graph section.

For preprocessing, another important work is candidate phrase generation. When extracting the dependency relations, a part-of-speech tags of the documents can be obtained. Therefore, any other pos tagger aren't needed to use for candidate phrase generation. Finally, candidate noun phrases are generated by matching sequential words with pattern *(adjective)*(noun)+*, which represents zero or more adjectives followed by one or more nouns.

4.2 Graph-based Term Ranking

After the dependency graph is constructed, the score associated with each vertex is set to an initial value of 1, and graph-based term ranking algorithm as described in equation 1 is run on the dependency graph until convergences. Once a final score for each vertex in the dependency graph is calculated, post processing phase would be run.

4.3 Learning anti-patterns

For processing, anti-patterns are learned from candidate noun phrases for training dataset. First, each candidate phrase for every document is assigned by a label that is representing whether the candidate phrase is keyphrase or not. The output of learning anti-patterns is the four lists of four kinds of words. Therefore, we first all possible words for the candidate phrases into the four lists. For example: if the candidate phrase has a three words, the first and last words would be added separately into the head and last lists, and all words would be added into the anti list. Otherwise, if the candidate phrase has only one word, this word would be added into the single list. In order to extract a useful patterns, the weak patterns should be removed by measuring the strength of patterns as described in equation 2. In equation 2, a value of threshold *c* is different for each kind list. The default value of threshold c for each kind of anti-pattern is shown in Table 1. if a support of the pattern doesn't satisfy threshold *c*, the pattern word would be removed from the list. Then a power(X) for every survived patterns

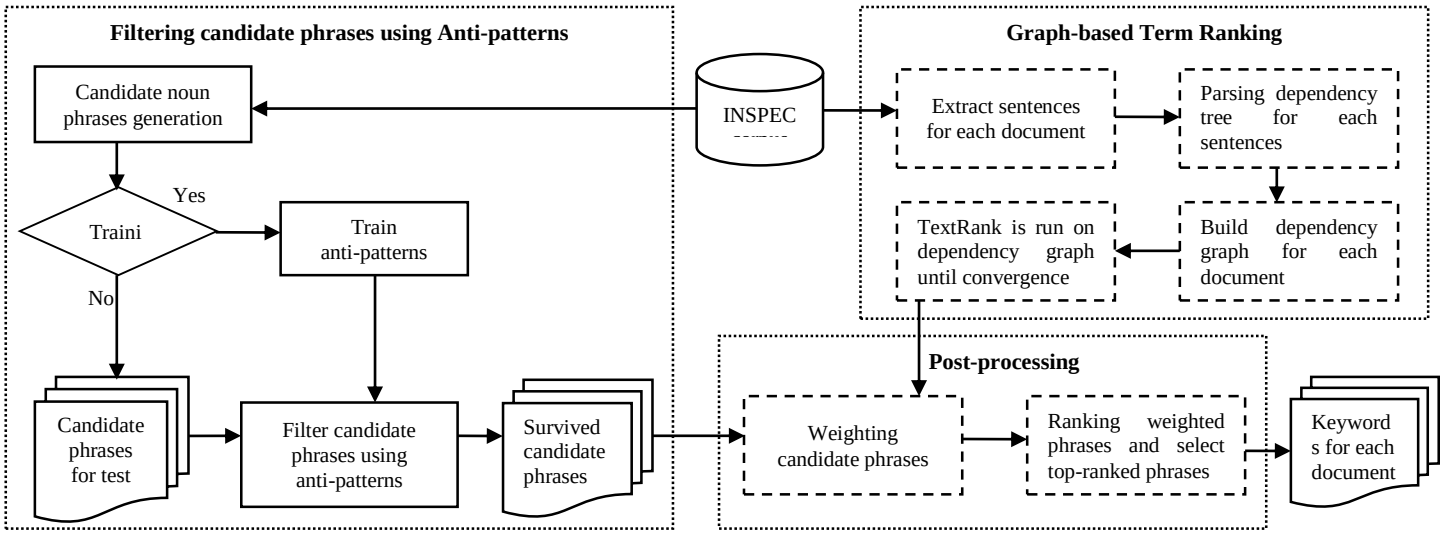


Figure 3. Flowchart of Proposed System

which satisfy threshold *c* is calculated using formula 2. Finally, the anti-patterns, which satisfy the condition Power(X) < 0.15, are survived and considered as strong anti-patterns.

4.4 Post-processing

Before starting post processing phase, all candidate phrases and anti-patterns are generated. Therefore, for this phase, all candidate phrases was filtered out by using anti-patterns. Then the survived candidate phrases are ranked as described in equation 3. The score of a candidate phrase pi is computed by summing importance scores of words contained in the phrase.

$$PhraseScore(X) = \sum_{v_j \in p_i} WordScore(v_j) \qquad (3)$$

All survived candidate phrases in the document are ranked in decreasing order of the phrases scores and top ranked k phrases are select as the keywords. k ranges from 1 to 25.

5. Experimental results and Evaluation

The Inspec corpus used in the experiments is a collection of 2000 abstracts with a paper title from journal papers for Computer Science and Information Technology. Each abstract has two sets of keyphrases assigned by the indexers: the controlled keyphrases which appear in the Inpsec thesaurus, and the uncontrolled keyphrases which do not necessarily appear in the thesaurus. The reason why we selected Inspec corpus is that is relatively the popular dataset for automatic keyphrase extraction, as it was first used by Hulth in [12] and later by Mihalcea and Tarau in [6] and Rafiqul in [14]. In her experiments, Hulth is using a total of 2000 abstracts, divided into 1000 for training, 500 for development, and 500 for test. Our system is based on the supervised learning. Therefore we used 1000 abstracts for training and 500 for test to compare previous existing systems. The evaluation was performed by comparing the system output to the human-annotated corpus in terms of the precision (p), recall (r) and their harmonic mean, the F-measure (F). These are based on the number of true positives (TP), false positives (FP) and false negative (FN) returned by the system

$$p = \frac{c_{correct}}{c_{extract}}, \quad r = \frac{c_{correct}}{c_{normal}}, \quad f = \frac{2pr}{p+r} \qquad (4)$$

where $c_{correct}$ is the total number of correct keyphrases extracted by a method, $c_{extract}$ the total number of automatic extracted keyphrases, and c_{normal} the total number of human-labeled standard keyphrases.

5.1 Influences of Anti-patterns

There are four kinds of anti-patterns including: a single word, head word, tail word, and anti-word. In Figure 5, the performances of each kind of anti-patterns are shown. Moreover, Figrue 5 shows the performances of the combination of all four kinds of anti-patterns and the baseline unsupervised method. From Figure 5, we can see clearly that when a number of keyphrases is smaller than 12, the f-measure of each method increases where the anti-word patterns beat other kinds patterns. However, when a number of keyphrases increases than 12, the f-measure of some method is dropping where single-word patterns is the best performing pattern. Because candidate phrases list contains a few phrases with one length if a number of keyphrases is small. Otherwise, a number of candidate phrase with one length increases dramatically. When a threshold parameters is set lower than the default values, the precision and recall of our method drops with together from the best result. Thus when a threshold parameters is set higher than the default values, recall of our method increases as precision and f-measures drop than the best result.

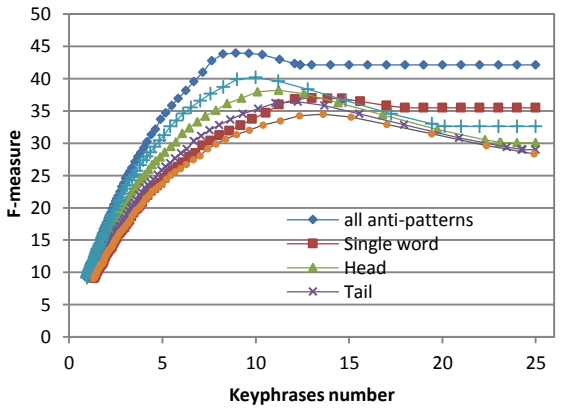


Figure 5. Influences of anti-patterns

5.2 Comparison between Dependency Graph and Word Co-occurrence Graph

In order to evaluate the efficiency of the dependency graph, this experimental work aims to demonstrate the difference between the word co-occurrence graph, the basic and modified dependency graphs. In Figure 6, we plots curves of each graph for both supervised and

Table 2. Comparing proposed method with other existing methods

Method	Assigned		Correct		Precision	Recall	F-measure
	Total	*Mean*	*Total*	*Mean*			
Textrank [6]	6,784	13.7	2,116	4.2	31.2	43.1	36.2
Hulth [12]	7,815	15.6	1,973	3.9	25.2	51.7	33.9
Rafiqul [14]	6114	12.23	2386	4.8	39.1	48.7	43.4
Unsupervised,EnglishPCFG	5,984	12.0	1,946	3.9	32.5	39.6	35.7
Supervised.EnglishPCFG	4,446	**8.9**	2,095	4.2	**47.0**	42.5	**44.7**

unsupervised methods. In general, the modified dependency graph is the best performing graph for both supervised and unsupervised methods. But when a number of candidate keyphrases is smaller than 5 for supervised method, the word co-occurrence graph beats the modified dependency graph. From Figure 6-A, we can see that it clearly outperforms other graphs for unsupervised method. After modifying the basic dependency graph, the performance of the keyword extraction method have been slightly improved. However, we prefer the modified dependency graph than the basic dependency graph. According to the results of Tfidf and Textrank on previous several studies in [18-19], these works proved that Tfidf is the best

performing unsupervised method although Tfidf is the simplest method. Moreover, the modified dependency graph exploits the main characteristics of Tfidf. Therefore, as shown in Figure 6-B, the modified

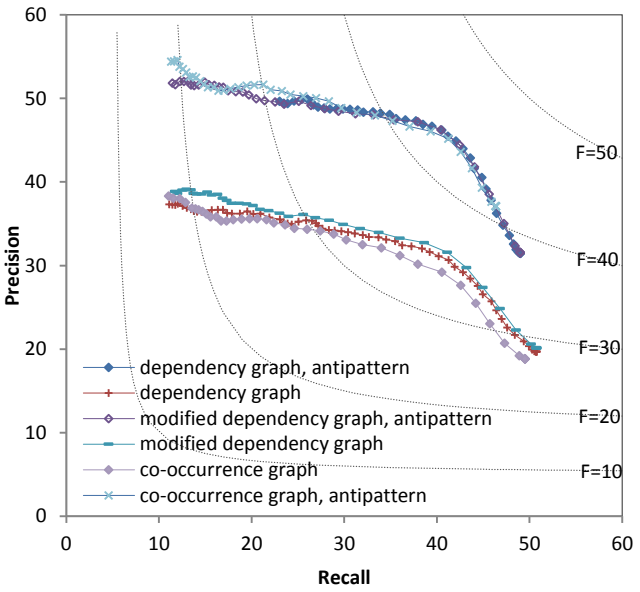


Figure 6. Comparison between Dependency graph and Co-occurrence graph

dependency graph is the graph with the best performance.

5.3 Comparing with other existing methods

In order to verify efficiency of a novel-graph based keyword extraction using anti-pattern, in Table 5, we compared its performances with a several previous works on the Inspec corpus. For each method, the table lists the total number of keywords assigned, the mean number of keywords per abstracts, total number of correct keywords, as evaluated against the set of keywords assigned by professional indexers, and the mean number of correct keywords. The table also lists precision, recall, and F1-measure.

When our method comparing with the best results of other existing methods on Inspec dataset, our system offered more few keyphrases than other systems. However, from Table 1, we can see that our novel graph-based keyphrase extraction method clearly outperformed other existing methods.

6. Conclusion

In this paper, we introduced the novel graph-based keyphrase ranking method using anti-patterns and showed how it can be successfully used. Our novel graph, called the dependency graph, is based on grammatical dependency relations between words. Furthermore, the important aspect of the dependency graph is that its grammatical relations between words are not limited by any window size. The anti-patterns are utilized to filter out some unwanted phrases from the candidate phrase list. Moreover, an important aspect of anti-pattern is that the strongest anti-patterns can help to filter out unwanted phrases in any domain, although the anti-patterns only covered on one specific domain. While the effect of each kind of anti-pattern was significant, these effects overlapped with each other when all kinds of anti-patterns were combined. The best balanced f-measure was the case where all kinds of anti-patterns were used. This suggests that the effect of each kind of the anti-pattern was small when all kinds of anti-patterns were combined. We have conducted the systematic evaluation of our proposed method. Several

conclusions can be drawn from our experimental works. Finally, the performance of our proposed method outperforms any other existing methods.

Acknowledgement

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA(National IT Industry Promotion Agency).

7. References

[1] Peter DT. "Learning Algorithms for Keyphrase Extraction", in *Journal of Information Retrieval*, 1999, pp 34-99.

[2] Kamal S, Mita N and Suranjan G. "Machine learning based Keyphrase Extraction: Comparing Decision Trees, Naive Bayes and Artifical Neural Networks", 2012, pp 693-712

[3] Wang J, Peng H, Hu J-S, "Automatic Keyphrase Extraction from Document Using Neural Network", *ICMLC 2005*, 2005, pp. 633-461

[4] Witten IH, Paynter GW, Frank E, Gutwin C, and Nevill-Manning CG. Kea: Practical automatic keyphrase extraction. In *Proc. of the 4th ACM Conference on Digital Libraries*, 1999.

[5] Medelyan O and Witten IH. "Thesaurus based automatic keyphrase indexing," in *Proceedings of the 6th ACM/IEEECS Joint Conference on Digital libraries*, ser. JCDL '06. New York, NY, USA: ACM, 2006, pp. 296–297.

[6] Mihalcea R and Tarau P. "Textrank: Bringing order into texts," in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, D. Lin and D. Wu, Eds. Barcelona, Spain: *Association for Computational Linguistics*, July 2004, pp.404–411.

[7] Wan X and Xiao J. "Single document keyphrase extraction using neighborhood knowledge," in *Proceedings of the 23rd National Conference on Artificial intelligence* - Volume 2, ser. AAAI'08. AAAI Press, 2008, pp. 855–860.

[8] Liu Z, Huang W, Zheng Y, and Sun M. "Automatic keyphrase extraction via topic decomposition," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Cambridge, MA: Association for Computational Linguistics, October 2010, pp. 366–376.

[9] Blei DM, Andrew YN, and Jordan MI, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[10] Zhao X, Jiang J, He J, Song Y, Achanauparp P, Lim EP et al. "Topical keyphrase extraction from twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, USA: Association for Computational Linguistics, June 2011, pp. 379–388.

[11] Abdelghani B and Mohammed A, "NE-Rank: A Novel Graph-based Keyphrase Extraction in Twitter" in *Proceedings of 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012, pp 372-379.

[12] Annette Hulth, "Improved automatic keyword extraction given more linguistic knowledge," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, M. Collins and M. Steedman, Eds., 2003, pp. 216–223.

[13] Thuy DN and Min-Yen K ,"Keyphrase Extraction in Scientific Publications", in *Proceedings of International Conference on Asian Digital Libraries (ICADL)*, 2007, pp. 317-326

[14] Md. Rafiqul I and Md. Rakibul I. "An Improved Keyword Extraction Method Using Graph Based Random Walk Model" In *Proceedings of 11th International Conference on Computer and Information Technology (ICCIT 2008)*, pp 225–229.

[15] Dan K and Christopher DM. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003, pp. 423-430.

[16] Dan K and Christopher DM. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. *In Advances in Neural Information Processing Systems 15 (NIPS 2002)*, Cambridge, MA: MIT Press, pp. 3-10.

[17] Marie-Catherine DM, Bill M and Christopher DM. 2006. "Generating Typed Dependency Parses from Phrase Structure Parses." In *LREC* 2006.

[18] Kazi SH and Vincent Ng. "Conundrums in Unsupervised Keyphrase Extraction: Making Sense of the State-of-the-Art" In *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp 365–373.

[19] Zede Z, Miao L, Lei Chen, Zhenxin Y, and Sheng C, "Combination of Unsupervised Keyphrase Extraction Algorithms" in *Proceedings of 2013 International Conference on Asian Language Processing*. 2013, pp 33-36.

Developing Graph Database for Multilingual Corpus

Hyeon Ah Park[1], Khuyagbaatar Batsuren[1], Nak Hyeon Choi[1], Jeong Hee Hwang[2]
[1]*Database/BioInformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*
{hapark, huygaa, nak} *@dblab.chungbuk.ac.kr*
[2]*Department of Computer science, Namseoul University, South Korea*
jhhwang@nsu.ac.kr

Abstract

The use of traditional parallel of multilingual corpora gives good contribution to better translation study and developing advanced machine translation technique – however, it still provides unclearness and confusion between different languages as these corpora relies on word-to-word compare and literal relationship provided from only one language. This paper describes of designing graph-based database corpus with basis of non-lingual representative of notions and object, annotated with words from various language instead of giving POS tags or other linguistic annotation to texts. We explain how this corpus can aid much flexible translation and be more helpful for multilingual translation study.

1. Introduction

Text corpus, a large structures data set of texts is used for computational linguistic research, often for developing better machine translation system, speech recognition and natural language processing [1]. Not just a database of words, many corpora are annotated with part-of-speech tagging or grammatical tagging to help analyze language by relationship between words and recognizing phrase and sentence, or lemma indication. Also there are not only monolingual corpora but as well as multilingual corpora too.

Corpora can be especially useful for translation study. A parallel corpus, which consists of texts originally written in a language A alongside with translation of language B, is good for providing links between source language and target language with alignment techniques, or information of language-pair specific translational behavior [2][3]. A multilingual corpus, unlike parallel corpus, is more like a set of un-translated original monolingual corpora but it could help to easily spot the practical problem of translation occur between different languages [4].

To combine these useful features of differently structured corpora, we suggest of designing a new way of creating multilingual corpora. We show how we can use graph-based database for creating a corpus more suitable for translation study involving more than two languages.

2. Multilingualism and Graph

When a multilingual speaker (including bilingual speaker) tries to translate a sentence, it can be seen that they are usually paraphrase instead of literally translating. This happens due to their knowledge of gap between simple word-to-word match and word-to-meaning match. For example, a verb going along with certain noun is shown to be presented in similar word pairs in various languages; ‘a man goes’ in English is translated ‘사람이 가다’ in Korean. The noun ‘man’ can be literally translated into noun ‘사람’ in Korean, and the verb ‘go’ and ‘가다’ are the same also. On the other hand, ‘Peel an apple’ and ‘Peel an orange’ is translated with different verbs in Korean despite English use the same verb; to smoothly translated them, each sentence should become ‘사과를 깎다’ and ‘오렌지를 까다’. In English, a verb ‘peel’ means ‘stripping off a surface’ while ‘깎다’ in Korean is more close to ‘cutting off a surface with a knife’ and ‘까다’ to be ‘taking off a shell from something’. Even when all these verbs are capable of indicating a behavior of eliminating skin from a fruit, we see that initial meaning can be quite different by not including same details depending on relationship with a specific noun. Translation with existing corpora can face

difficulty in connecting more plausible verbs or adjectives with a noun around multiple languages, because it depends on grammatical tagging of a word but not the actual detailed action or situation. Using frequency of certain noun-verb matching or adjective-noun matching could fail also, as sometimes even noun isn't the same literal translation of one another.

To solve such problem, we present a design of multilingual corpus which does not use any language for the source or the basis for language comparison. Instead, we first construct a graph database representing relationship between objects and notions expressed with ID with no linguistic meaning as nodes, and tagging words from different languages. Instead of giving POS tags to a word, we distinguish nodes by grammatical position at the first place so each words from different languages can naturally gathered by their position in a sentence. A graph can successfully capture and map of any kind of relationship between words in any position if proper nodes distinguisher is given. This "relationship" represented in edges inside graph is able to map any structure of a sentence and even provide delicate differences between various languages.

3. Reverse Tagging

The construction of our corpus database and idea of reverse tagging in described in Figure 1 and Figure 2.

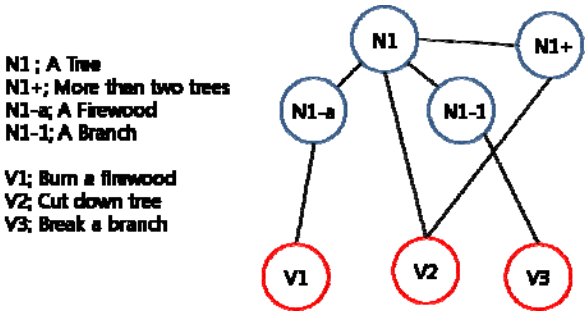


Figure 1. Graph-based corpus based on non-lingual indicator nodes

Notice that in Figure 1 singular noun and plural noun are clearly separated and verbs are meant to be representing more specific use of an action instead of being a lemma. Nouns having relationship indicates they are closely related objects thus sometimes share a same verb. The level of being general or much detailed idea of an object and action along with verbs in different tense can be given by adding nodes and proper IDs. Developing efficient ID system for such corpus will be studied further in the future.

Multiple annotations of words from different languages are provided for each node. Flexible translation is available when a sentence '나무를 태우다' is given, for example, literally this can be simply translated as 'burn a tree', but if the original Korean text is meant to be N1-a and V1 relationship, it can also return the sentence 'burn a firewood' as well.

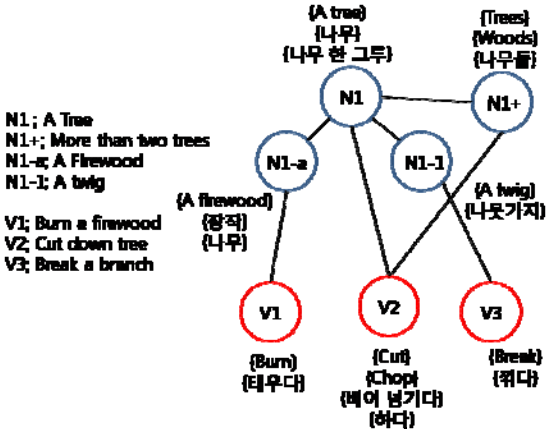


Figure 2. The reverse tagging

Because '나무' itself is too broad to mean 'firewood' simple alignment cannot easily capture the alternative translation. If postposition or preposition is added, it could achieve more clear and explicit translation. The more nodes are created, the better translation becomes possible, because it will create more detailed relationship between notions and object.

As mentioned earlier, this graph-based corpus can capture both information of translational behavior of different languages by letting researchers to analyze which words are gathered along with which nodes of detailed concepts and the picture of how various languages draw divergent map when trying to describe similar objects and ideas.

4. Conclusion and Further Research

As traditional parallel or multilingual corpora could show limited efficiency in aiding translation, we have suggested a new design of corpus which is graph-based with non-linguistic ID annotated with actual words from different languages. Such design of corpus can promise new efficient way of multilingual translation study and flexible translation but there are problems to be noted.

First, to make this corpus to return better translation a good ID system must be developed for creating nodes with very much detailed meanings to capture

even small gap between words sharing similar meanings, but still not exactly the same.

Second, for the actual translation an algorithm to analyze source language, recognizing the context, creating new sentence with target language from the context and corpora match should be developed.

Third, if not adding nodes and annotations manually by multilingual speakers, another computational text analyzer and annotation method for automatically picking up any possible “hidden” contextual meaning of words have to be developed.

With these three problems we will continue to enrich the graph-based corpora and perform experiments of testing the accuracy of translation outcome with sample texts in the future.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2008-0062611).

5. Reference

[1] Yoon, H., & Hirvela, A, “ESL student attitudes toward corpus use in L2 writing”, *Journal of second language writing*, Vol.13, No.4, 2004, pp. 257-283.

[2] Kenny, D, *Corpora in translation studies*. Routledge encyclopedia of translation studies, 1998.

[3] Mariani, E., Peters, C. and E.Picchi, "Bilingual Reference Corpora: Creation, Querying, Applications", *Ferene Kiefer, Gabor Kiss and Julia Pajzs (Eds),Papers in Computational Lexicography: Complex 92*, Budapest, Linguistics Institute, Hungarian Academy of Sciences, 1992, pp. 221-228.

[4] Aijmer, K., Altenberg, B., & Johansson, M, “Languages in contrast”, *A symposium on text-based cross-linguistic studies*, Lund studies in English, Vol.88, 4-5 March 1994.

Classification of Diseases from Number of Outbreaks

Wooyeong Jang[1], Changwoo Woo[1], Ho Sun Shon[1], Young-Sung Lee[1], Young Gyu Kim[1], Keun Ho Ryu[2]

[1]*Graduate School of Health Science Business Convergence, Chungbuk National University, South Korea*

{*jangwy8838, cwwoo.psm, shon0621, lee.medric, brsurg*}*@gmail.com*

[2]*Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*

khryu@dblab.chungbuk.ac.kr

Abstract

As the demographics of society shift towards a more aging society and national income increases, various medical expenses influence the medical service industry. With this social structure and changing environment, actions for new diseases are needed urgently. However, to date there is only a disease prediction model for specific diagnosis, not a generalized prediction model on the national level.

This paper suggests methods for classification of diseases and interprets the distribution of classified diseases by position where diseases come from visually.

There are 4 generalized types of diseases. These are environmental diseases, lifestyle related disease, mental illness, endemic diseases.

Through this classification, this paper makes developing and evaluating a disease prediction model possible. Also, this result could contribute to the evidence base for public health and healthcare policy decision making.

1. Introduction

Since 2000, due to increasing national income, the needs of various customers have started surfacing in the medical service industry. Also Medical treatment demand has been changing from major diseases to new diseases because of the increasing aging society, along with increasing stress from rapidly change societal factors and food life. As these changes occur, we need action for new diseases. There is no generalized prediction model, only a diffusion model. A risk model and prediction model for specially fixed diagnosis is developing nationally [1].

Also there is a health *belief model* generally related to diseases. However, this model does not influence enough the prediction of diseases [2].

The Korean Standard Classification of Diseases and Causes of Death are used to classify diseases and other health problems recorded on many types of health and vital records including death certificates and health records. In addition to enabling the storage and retrieval of diagnostic information for clinical, epidemiological and quality purposes, these records also provide the basis for the compilation of mortality and morbidity statistics[3].

Administrative district of Korea consists of 1 metropolitan city, 6 megalopolis, 8 *provinces*, 1 Special Autonomic City, and 1 Special Autonomic *province* which uses sovereign power. The gross 17 administrative districts are classified as metropolitan municipalities [4].

2. Subject

2.1. Collecting data

In this study, gross data from a 36 month period from January 2010 to December 2012 was used. The classification of diseases, three-stage diagnosis, is 1725.

There were 16 administrative districts except for Sejong-si during the period of data collection. But there are 17 fiducially administrated districts for 2014.

Table 1. Information from Collected Data

Collecting data time (month)	Classification of Diseases	Separate cities and provinces
2010.1 ~ 2012.12	third-degree diagnosis	Administrative District
36	1,725	16

Table 2. Classification of the graph is defined by four types of diseases from basic theory

List	Diagnosis and Definition	
Environmental disease	J30	Other diseases of upper respiratory tract
	J45	Asthma
	L20	Atopic dermatitis
Lifestyle related disease	I10-I15	Hypertensive diseases
	E10-E14	Diabetes mellitus
Mental illness	F31	Manic episode
	F32	Bipolar affective disorder
	F33	Depressive episode
Endemic disease	B50-B54	malaria

2.2. Classification of diseases

Before developing a prediction model, the diseases should be defined using the available model. Diseases in the graph are classified based on the four main types that come from basic theory. There are 3 typical environmental diseases such as diseases of the upper respiratory tract, Asthma, and Atopic dermatitis. In diseases related to lifestyle, there are diseases of the upper respiratory tract. In mental illness, there are diseases of manic episodes, bipolar affective disorder and depressive episodes. In endemic diseases, there are diseases of malaria [5].

Many other diseases are included in the 4 types. Several specific diseases which are not available in the upper part are included in other types. The prediction model should be started from these types.

2.3. Occurrence distribution of patients with the disease

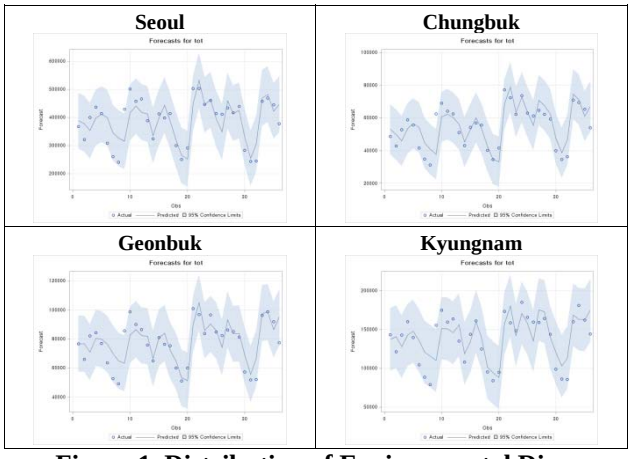


Figure 1. Distribution of Environmental Disease

Occurrence for patients of lifestyle-related disease is slightly different from occurrence for seasonal patients and normally has a wide confidence interval.

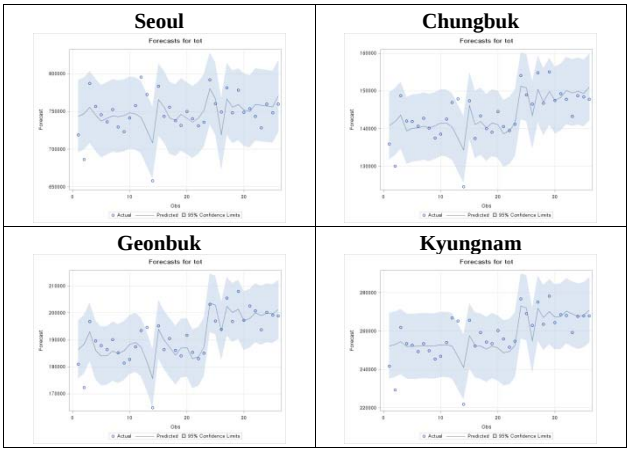


Figure 2. Distribution of Lifestyle related Disease

Occurrence for patients of environmental diseases changes seasonally and normally has a short confidence interval.

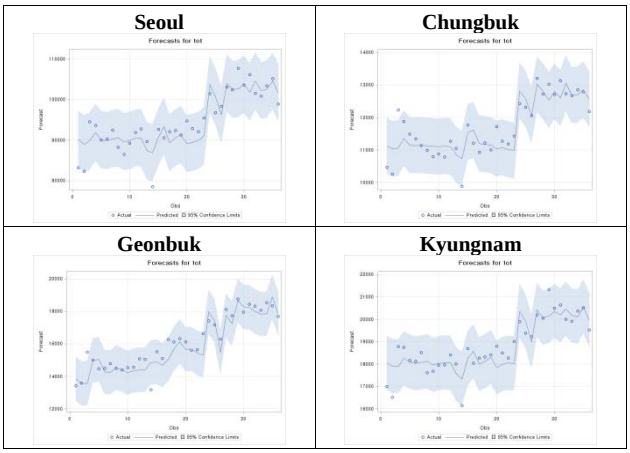


Figure 3. Distribution of Mental Illness

Distribution of mental illness is similar in each area but both environmental diseases and lifestyle related diseases have a different distribution.

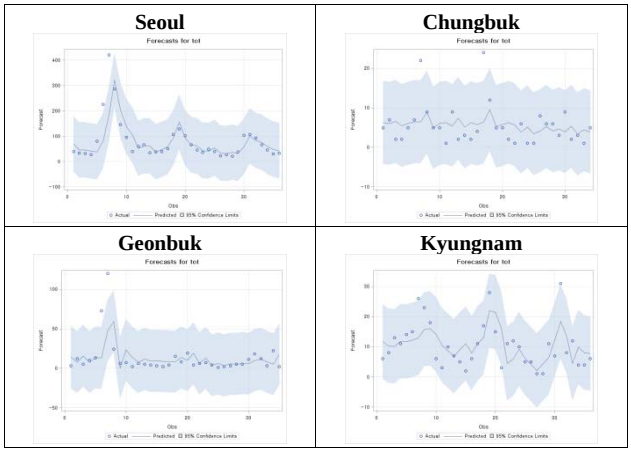


Figure 4. Distribution of Endemic Disease

Endemic diseases are have a different occurrence distribution depending on the region.

3. Conclusion and Future Work

Diseases are defined as four types from basic theory. In environmental diseases, occurrence points of patients are seasonal and have a short confidence interval. In lifestyle-related diseases, occurrence points of patients are few seasonally and have a wide confidence interval. The distribution for patients of regional diseases varies greatly. We could develop a general prediction model using the occurrence distribution of diseases. In a general prediction model, health behavior and health condition will be added as variables for internal factors. For other factors, weather and national policy will be added.

Acknowledge

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), ISBB (International Science Business Belt) support program (2013K001552), and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No-2013R1A1A206518).

4. References

[1] Park Seong-Hoon**,** "Strateges for u-Health", *Gyeonggi Research Institute*, 2009, pp.1-4.

[2] JENNIFER A. HANSON, MS; JAMIE A., "Use of the Health Belief Model to Examine Older Adults' Food-Handling Behaviors", *Journal of Nutrition Behavior*, 2002, Vol 34, pp.25-30.

[3] Korean Classification of Diseases(KCD), *National Statistical Office*, 2004.

[4] Classification of administrative district, *National Statistical Office*, 2014.

[5] http://www.hira.or.kr/main.do

[6] Akaike, H., "A New Look at the Statistical Model Identification", *IEEE Transaction on Automatic Control*, AC–19, 1974, 716–723.

[7] Anderson, T. W., The Statistical Analysis of Time Series, *New York: John Wiley & Sons*, 1971.

[8] Andrews and Herzberg , A Collection of Problems from Many Fields for the Student and Research Worker, *New York: Springer–Verlag*, 1985.

Development of Web-based System for Analysis of Urinary Cancer Patient from Disease Prevention Questionnaire

Kyeong Seok Lee[1], Hyun Woo Park[1], Soo Ho Park[1], Kyung Ah Kim[2]
[1]*Database/Bioinformatics Laboratory, Department of Computer Science, Chungbuk National University, South Korea*
{kslee, hwpark, soohopark }@dblab.chungbuk.ac.kr
[2]*Department of Biomedical Engineering, School of Medicine, Chungbuk National University, South Korea*
kimka@chungbuk.ac.kr

Abstract

This paper constructs a nutrient database and implements system for analysis of urinary cancer patients. The proposed system stores diet, eating habits and lifestyles from questionnaire, and a web-based system calculates nutrient to manage the patients. This system would be helpful to analyze the relationship of urinary cancer, nutrient, eating hobby and lifestyle. Also our research is expected that can be meaningful research for prediction of the prognosis of the urinary cancer patients.

1. Introduction

In recent years, there have been many researches for relevance of cancer prevention, diet and lifestyle. The relationship between cancer and diet is closely related, it is proven much throughout many studies.

Dietary fiber was found to have effect to prevent colon cancer and rectal cancer [1]. Beta carotene has efficacy of prevention of gastric and lung cancer [2]. Vitamin E has efficacy of prevention of lung cancer and prostate cancer [3]. Vitamin C is effective in preventing lung cancer and gastric cancer [4]. However, the results of many researches are not consistent with relevance between cancer and diet, and we don't know which nutrients effective in the prevention of cancer [5]. Therefore we need to analyze the relationship between cancer and diet for prediction of prognosis of cancer.

In this paper, we construct a database for history of disease, diet and patients lifestyle. And we propose a web-based system for analysis of urinary cancer patients from the disease prevention questionnaire. The proposed system is expected that manage the patients data and analyzes the relationship of urinary cancer, nutrient, eating hobby and their lifestyle efficiently.

2. Related Work

Han et al. developed computerized nutrition counseling system for patients with diabetes [6]. In this system, they designed to find out a personal dietary history and to give suggestions about his incorrect dietary habit. And they analyzed the energy and nutrients of food consumed.

Han purposed system for nutritional assessment and diagnosis of dietary intakes through internet [7]. This study assessed the general status of the body such as ideal body weight, an obesity index, basal metabolic rate and total energy requirement. And they analyzed energy and nutrients of dietary intake including intake of dietary fatty acids and evaluated the nutritional status of dietary intake by comparing the energy and nutrient intake with recommended dietary allowance.

Son et al. proposed a system for analyzing nutrition status based on hierarchical fuzzy inference approach [8]. In this system, they analyzed the transition process on the nutritional status from an obesity degree, the previous nutritional status, and the eating pattern with an individual. And they evaluated obesity degree and final nutrition status.

3. Web-based Disease Prevention Question-naire Management System

This system framework is as shown Figure 1. Each patient's data investigated from disease prevention questionnaire, answers to each question were stored in

nutrient database. After selecting data of patients, each data was loaded in nutrient database and calculated total nutrients of each data. Finally, users can confirm the each amount of nutrients like energy, protein, fat, carbohydrate, Ca, vitamin, and so on. This overall system was developed using Flex 4.6 and JSP, and database was constructed using MySQL.

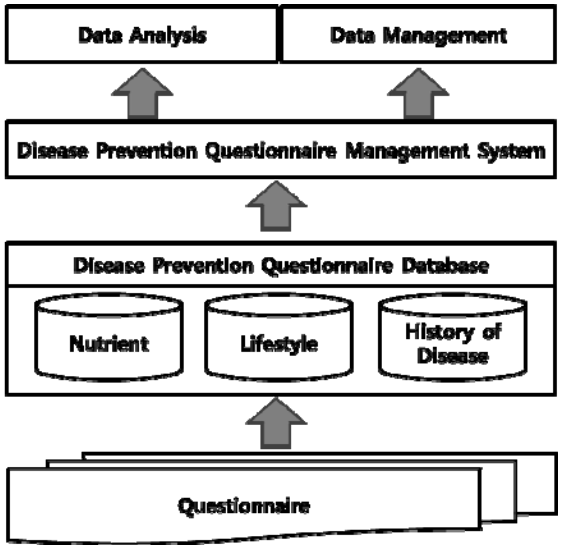


Figure 1. The framework for web-based disease prevention questionnaire management system

Disease Prevention Questionnaire Database

We collected data using questionnaire from the hospital. This questionnaire consists of 28 pages, and it included basic data of patients, history of medical, history of family, details of smoking, drinking, sleeping,
physical activities and dietary intake.

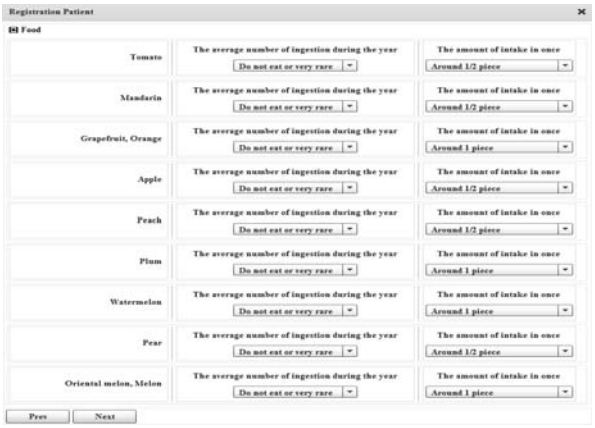


Figure 2. Nutrient, lifestyle, history of disease data collection from questionnaire

Nutrient, lifestyle, history of disease data collection from questionnaire is as shown Figure 2. It was designed to select or write answer about patients. After answered all questions, they can store data about patients in disease prevention questionnaire database.

Database design is as shown Figure 3. It was designed based on the questionnaire, answers of question were stored in each table. This database consists of 11 tables such as history of use of drug, history of family, food, job, medical, operation, patient, sleep, smoking and standard.

Each table included data of answer about detailed question of each category. For example, food table included data of answer about amount and number of intake of each food for a year. Other table included data of answer like food table. But standard table has nutrients weight degree on each food and question. When calculate nutrients, this table is used.

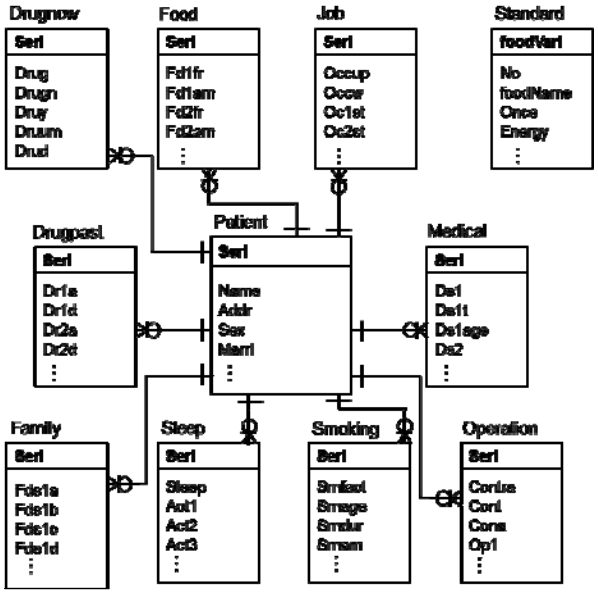
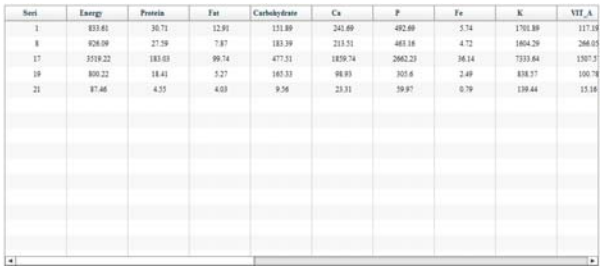


Figure 3. Disease prevention questionnaire database design

Calculation of patient's annual nutrient intake

This system was designed to calculate patient's annual nutrient intake. We imported data of answer of selected patient from disease prevention questionnaire database. Data of answer included amount and number of intake of each food for a year. We calculate amount of intake of each food using standard table. It calculated the product of amount and number of intake and weight degree which each nutrients. And it

summed amount of intake each nutrient of all food. Finally, we can confirm calculation of patient's annual nutrient intake. Calculation of patient's annual nutrient intake is as shown Figure 4. Association of cancer and nutrients can be analyzed using calculated amount of nutrient intake.



Seri	Energy	Protein	Fat	Carbohydrate	Ca	P	Fe	K	VIT_A
1	833.61	30.71	12.91	151.89	241.69	492.69	5.74	1701.89	117.19
8	926.09	27.59	7.87	183.39	213.51	483.18	4.72	1804.29	266.05
17	3519.22	183.03	99.74	477.51	1859.74	2662.23	36.14	7333.64	1507.5
19	800.22	18.41	5.27	165.33	98.93	303.6	2.49	838.57	100.78
21	87.46	4.35	4.03	9.56	23.31	59.87	0.79	139.44	15.16

Figure 4. Calculation of patient's annual nutrient intake

4. Conclusion and Future Directions

In this paper we constructed database for history of disease, nutrients and patients lifestyle from disease prevention questionnaire. And we developed the web-based system which analyzes relevance between cancer and nutrients to predict prognosis and prevention of cancer. This system can manage the patient's questionnaire data and calculate their annual intake nutrients. It can be expected to help analysis of relevance between cancer and nutrients. Currently, we stored to database only questionnaire data such as disease history, food, job, sleeping and smoking. In the future, it is possible to integrate the cancer patient's clinical database and provide various analysis methods for comprehensive analysis of cancer patients.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2008-0062611).

5. References

[1] M.C.J.F. Jansen, H.B. Bueno-de-Mesquita, R. Buzina, F. Fidanza, A. Menotti, H. Blackburn, A.M. Nissinen, F.J. Kok and D. Kromhout, "Dietary fiber and plant foods in relation to colorectal cancer mortality", *The Seven Countries Study. International Journal of Cancer*, Vol 8(2), 1999, pp. 174-179.

[2] G. van Poppel and R.A. Goldbohm, "Epidemiologic evidence for beta-carotene and cancer prevention", *The American Journal of Clinical Nutrition*, Vol 62(6), 1995, pp. 1393-1402.

[3] J.M. Chan, M.J. Stampfer, J. Ma, Rimm E.B., Willett W.C. and Giocannucci E.L., "Supplemental vitamin E intake and prostate cancer risk in a large cohort of men in the United State", *Cancer Epidemiol Biomarker Prev*, Vol 9, 1999, pp. 893-899.

[4] Voorrips L.E., Goldbohm R.A., Brants H.A., van Poppel G.A., Sturmans F, Hermus R.J. and van den Brandt P.A., "A prospective cohort study on antioxidant and folate intake and male lung cancer risk", *Cancer Epidemiol Biomarkers Prev*, Vol 9, 2000, pp. 357-367.

[5] M.S. Lee, I.P. Chung and J.J. Jang, "Cancer Prevention and Diet", *Journal of Korea Association of Cancer Prevention*, Vol 7(2), 2002, pp. 210-214.

[6] J.S. Han and S.H. Rhee, "A computerized nutirition counseling system for patients with diabetes", *Journal of the Korean Society of Food Science and Nutrition*, Vol 22(6), 1993, pp. 734-742.

[7] J.S. Han, "A system for nutritional assessment and diagnosis of dietary intakes through internet", *Journal of the Korean Society of Food Science and Nutrition*, Vol 29(6), 2000, pp. 1177-1184.

[8] C.S. Son and G.B. Jeong, "A nutrition status analysis system based on hierarchical fuzzy inference approach", *Korean institute of intelligent systems*, Vol 17(6), 2007, pp. 731-737.

Session : Communication & Signal Processing

- Determination of Surface Radio Refractivity over Mongolia
  Jamiyan Sukhbaatar, Nyamjav Jambaljav, Damdinsuren Erkhembayar
- Constructing a System for Monitoring, Managing Groundwater in the Industrial Zones of Hanoi City
  Vu Thi Hong Nhan
- Investigation of SEE on a 32-bit Microprocessor based on SPARC V8 Architecture by Laser Test
  Chunqing Yu

Determination of Surface Radio Refractivity over Mongolia

Jamiyan Sukhbaatar, Nyamjav Jambaljav, Damdinsuren Erkhembayar
Department of Electronics and Communication Engineering,
School of Applied Sciences and Engineering,
National University of Mongolia
{jamiyan, nyamjav, eds}@num.edu.mn

Abstract

The seasonal and diurnal variation of radio refractivity over Mongolia was studied. The values of radio refractivity have been determined in Ulgii, Khatgal, Bayan-Ovoo, Baitag, Dalanzadgad, Bayandelger and Ulaanbaatar. The seasonal and diurnal refractivity was calculated for the period of thirty years meteorological data from 1984 through 2013. A total of more than one million refractivity measurements was considered in this analysis. The results indicate that the radio refractivity of a seasonal variation with high value in the wet season and low value in the dry season. This is a result of variations in meteorological parameters such as humidity, temperature and atmospheric pressure. The yearly maximum radio refractivity, 328 N-units was in Khatgal in the January and minimum one, 295 N-units was in Dalanzadgad in the May

Keywords : *Radio refractivity; radio-wave propagation; atmospheric propagateion; meteorological parameters*

1. Introduction

The radio refractive index n of the troposphere is an important factor in predicting performance of terrestrial radio links. Near the surface of the earth radio refractive index, n, is a number of the order of 1.0003 [1]. The radio refractive index in the troposphere is affected by the variations of meteorological parameters such as temperature, pressure and relative humidity. The changes in the value of the radio refractive index can curve the path of the radio wave. Even small changes in these variables can make a significant influence because radio signals can be refracted over the whole of the signal path [2].

Radio refractivity N is used in order to notice the changes in the values of the refractive index which is usually small. These N is obtained by subtracting 1 from the refractive index and multiplying the remainder from the refractive index and multiplying the remainder obtains units by a milling ($N=(n-1)*10^6$) [3]. In this way more manageable numbers are obtained.

Change in refractive index with height causes radiowaves to curve downwards, and to a degree which depends on the vertical refractivity gradient. Refractive bending causes extension of the radio horizon beyond the optical horizon. Surface radio refractivity N_s is known to have a high correlation with radio field strength values [4] while the surface refractivity gradient which depends on NS determines the refractive condition of the atmosphere which may result in a normal, sub-refractive, super-refractive or ducting layer, each of which has important influences on propagation of VHF, UHF and microwaves in the atmosphere. Under normal atmospheric conditions the refractive index of air decreases uniformly with height, and the surface value N_s is known to have a good positive correlation with the parameter ΔN representing the refractivity gradient in the first 1 kilometer above the surface. Thus, good knowledge of N_s is practically useful in planning and design of microwave communications systems. This work is aimed to find out the diurnal and seasonal variation of the surface radio refractivity over Mongolia.

Mongolia is a landlocked country in Northeast Asia located between the latitudes of 41°35'N and 52°09'N and the longitudes of 87°44'E and 119°56'E. Mongolia's territory reaches relatively high altitudes: while the average altitude is 1580 meters above sea level, 81.2% of the territory is higher than 1000 meters, and half of the territory is higher than 1,500 meters. In Mongolia, all natural zones such as high mountains, valleys between the mountain ranges, wide

steppe, desert and semi-desert zones are combined. Ecologically, Mongolia occupies a critical transition zone in Central Asia: here the great Siberian taiga forest, the Central Asian steppe, the high Altai mountains and the Gobi desert converge. The northwest and central parts of Mongolia are high mountainous regions, while the eastern part is a vast steppe region. The southern part of the country represents the semi-desert and desert area that is known as Mongolian Gobi [5].

Since Mongolia has many types of landscapes, in this study, we investigated the diurnal and seasonal variation of surface refractivity over seven localities (Ulgii, Khatgal, Bayan-Ovoo, Baitag, Dalanzadgad, Bayandelger and Ulaanbaatar) to cover all the climatic regions in Mongolia. Meanwhile, this study covers a period of thirty years' meteorological data from 1984 through 2013.

2. Climatic characteristics of Mongolia and Data Acquisition

The climate of Mongolia is a harsh continental climate with four distinctive seasons, high annual and diurnal temperature fluctuations, and low rainfall. Average annual temperatures range between 8.5°C in the Gobi and -7.8°C in the high mountainous areas. The extreme minimum temperature is usually between -31.1°C and -52.9°C in January and the extreme maximum temperature ranges from +28.5°C to +42.2°C in July. The average annual precipitation is low (200-220 mm) and represents a range between 38.4mm per year in the extreme South (Gobi desert region) and 389 mm per year in limited areas in the North. Most precipitation occurs in the months of June, July and August; the driest months occur between November and March.

Seven weather stations (located at Ulgii, Khatgal, Bayan-Ovoo, Baitag, Dalanzadgad, Bayandelger and Ulaanbaatar) which studied in this work were selected from six regions according to their landscape characteristics. First weather station at Ulgii is located in coldest region, which includes the Lakes Basin. Second weather station at Khatgal is located in wettest region, which includes the Khangai and Khuvsgul mountains and arable land areas. Third weather station at Bayan-Ovoo is located in a region where receives a moderate amount of precipitation and includes the steppe and the Khentei mountains. Fourth weather station at Baitag is located in the driest and hottest climate region, covering the Altain-Gobi and Gobi-Altain mountains. Fifth weather station at Dalanzadgad is located in the Gobian zone. Sixth weather station at Bayandelger is located in the Dornod-Gobian zone. Seventh weather station is located in Ulaanbaatar, capital city of Mongolia. Latitude, longitude, altitude, station number and locality of the weather stations are presented in Table 1.

Table 1. Latitude and longitude of the stations

Station Number	Localities	Latitude (North)	Longitude (East)	Altitude	Climatic region
44214	Ulgii	48.93°	89.93°	1715	Coldest region, includes the Lakes Basin
44207	Khatgal	50.43°	100.15°	1668	Wettest region, includes Mountains area
44302	BayanOvoo	47.78°	112.11°	926	Steppe and Khentei mountains
44265	Baitag	46.11°	91.46°	1186	Hottest region, covering Altain-Gobi
44373	Dalanzadgad	43.58°	104.41°	1465	Gobian zone
44352	Bayandelger	45.73°	112.36°	1101	Dornod-Gobian zone
44292	Ulaanbaatar	47.92°	106,84°	1306	Capital city

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:

$$N = (n-1)\times 10^6 \tag{1}$$

Radio refractivity [3] N is expressed by:

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

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

$$N_{dry} = 77.6\frac{P}{T} \tag{3}$$

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

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

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]:

$$e = \frac{He_s}{100} \tag{5}$$

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

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

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

The values of the refractivity N have been determined by using (2). Thirty years (1984-2013) values of temperature, humidity and atmospheric pressure were taken from seven weather stations (see Table 1). Dry term, N_{dry}, and wet term, N_{wet}, of radio refractivity were determined by using (3) and (4) respectively. The partial water vapor pressure e was determined by using (5) and (6). All calculations have been performed using *Mathematica* [6].

The seasonal variation of dry term, N_{dry}, wet term, N_{wet} and radio refractivity N in Ulgii, Khatgal, Bayan-Ovoo, Baitag, Dalanzadgad, Bayandelger and Ulaanbaatar are presented in Figures 1-3. The result showed an increase in the value of refractivity from a minimum value of about 295 N-units at Dalanzadgad station to a maximum value of about 328 N-units at Khatgal station. It is observed from the results that refractivity values at Dalanzadgad station are lower than other stations. Dalanzadgad is located on Gobian zone and is dominated by dry and hot season within the year. The result also showed that higher values of refractivity are in December, January, July and August. Most precipitation occurs in these months in Mongolia. Refracvity dropped in April, May, September and October which are the driest months of the year.

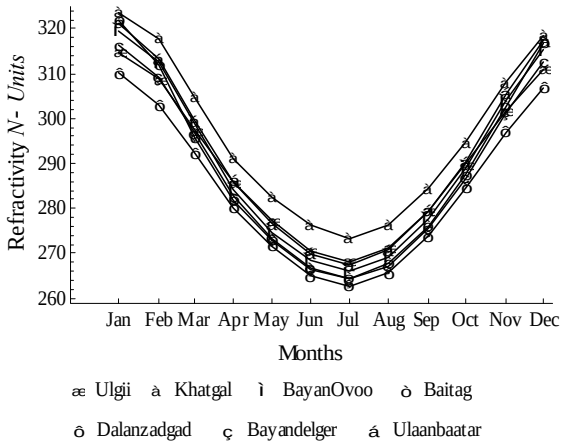


Figure 1. Seasonal variation of dry term

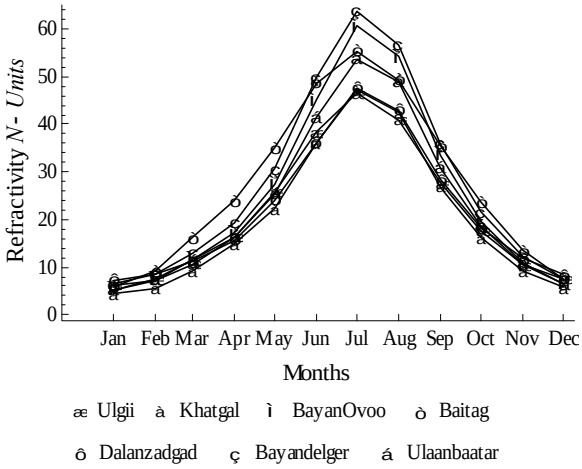


Figure 2. Seasonal variation of wet term

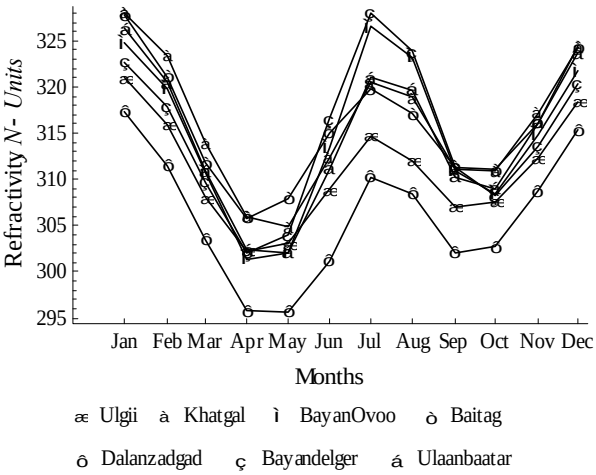


Figure 3. Seasonal variation of radio refractivity

Temperature, humidity and atmospheric pressure, which are main meteorological parameters used for calculation of radio refractivity in seven localities are presented in Figures 4-6.

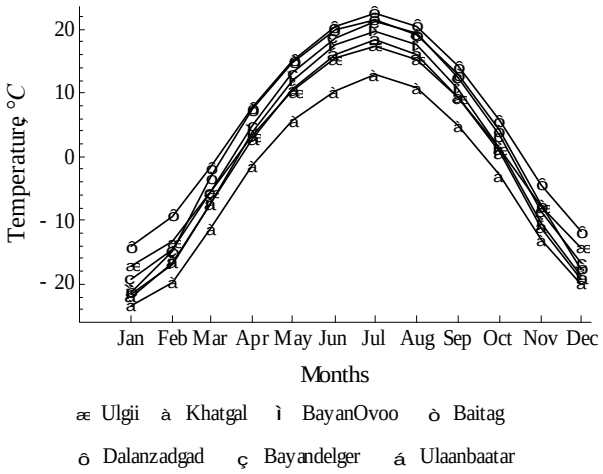


Figure 4. Seasonal variation of temperature

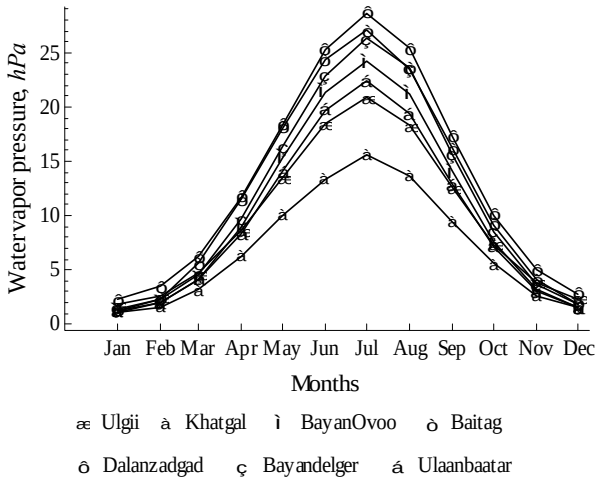


Figure 5. Seasonal variation water vapor pressure

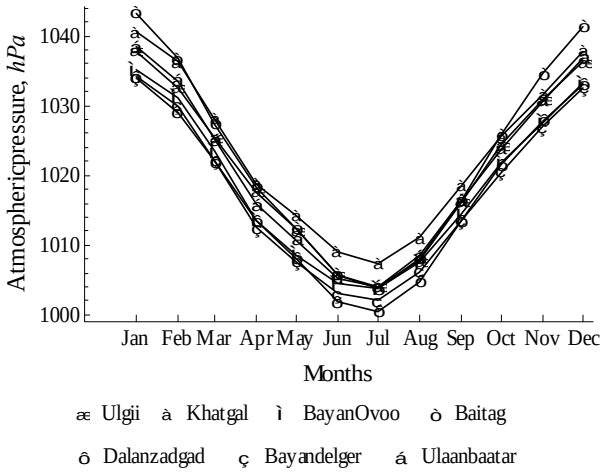


Figure 6. Seasonal variation of atmospheric pressure

The diurnal variations of refractivity in different places at Ulgii, Khatgal, BayanOvoo, Baitag, Dalanzadgad and Bayandelger and different months are depicted in Figure 7 to Figure 10 respectively. Each day, eight measurements of temperature, relative humidity and pressure were taken at 0.00, 03.00, 06.00, 09.00, 12.00, 15.00, 18.00 and 21.00 hours local time at all six stations.

The diurnal variations of radio refractivity at six localities in January are shown in Figure 7. January is the coldest month in Mongolia. From the result, the value of refractivity reached a minimum around 09.00 to 08.00 hours local time and a maximum value around 21.00 to 0.00 hours local time for all localities. The maximum value of radio refracvity about 332 N-units was in Khatgal around 0.00 hours and the least values 312 N-units was in Dalanzadgad around 09.00 hours.

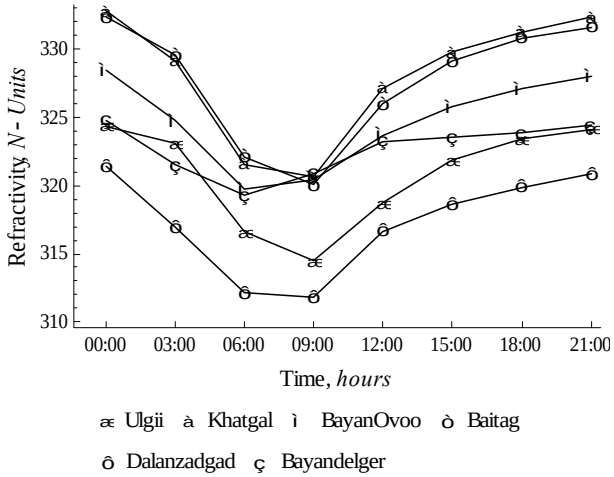


Figure 7. Diurnal variaion of radio refractivity for january

The dependencies of diurnal values of N on the time of day in different places in April are presented in Figure 8. The driest month occurs in April. Figure 8 shows that the refractivity is high (about 312 N-units in Khatgal and Baitag) in the midnight (around 0.00 hours) for all localities. It gradually drops from 0.00 hour reach a minimum of around morning and gradually risen till the end of the day. The minimum value (291) of refractivity was in Dalanzadgad.

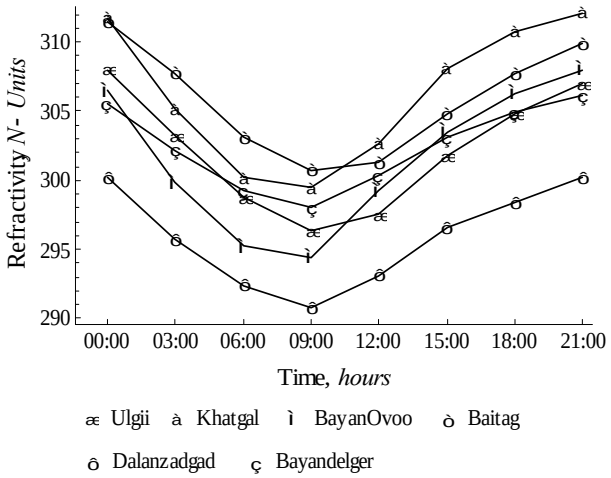


Figure 8. Diurnal variation of radio refractivity for april

Figure 9 and Figure 10 representing the diurnal variations of radio refractivity of July, the hottest month and August, the wettest month of the year respectively. Most precipitation occurs in August. The results showed that the variations of refractivity values in July and August are almost same for all regions. A maximum value of about 331 N-units was observed about 21.00 hours over BayanOvoo and Bayandelger in July. In July and August, minimum diurnal variations (302 N-units) were observed over Dalanzadgad at 09.00 in August.

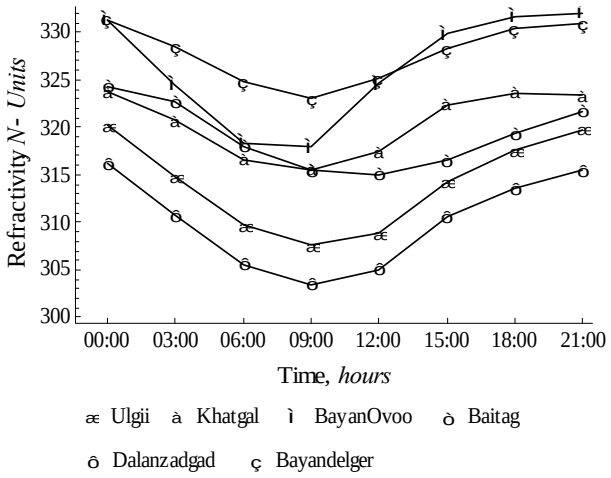


Figure 9. Diurnal variation of radio refractivity for july

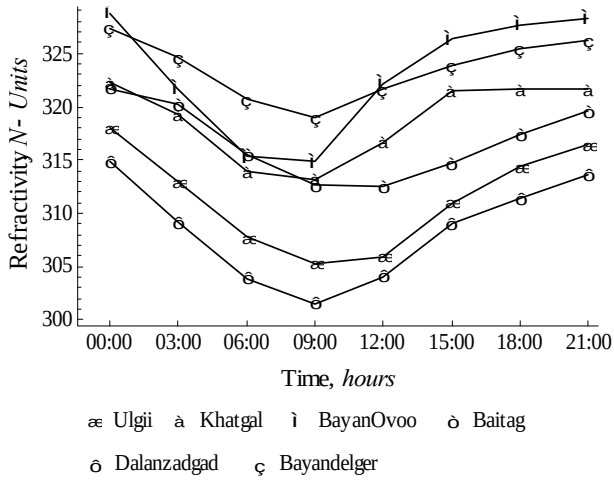


Figure 10. Diurnal variation of radio refractivity for august

Comparison of Diurnal variation of radio refractivity with time in January, April, July and August in Ulaanbaatar is depicted in Figure 11. The hourly variations of meteorological parameters for each day for the thirty minute interval for each day were recorded in Ulaanbaatar. In Figure 11, diurnal average values of refractivity in January are higher than other months and the diurnal variations of N-values are smaller in April. The variations in the daily N-values was almost same pattern for months.

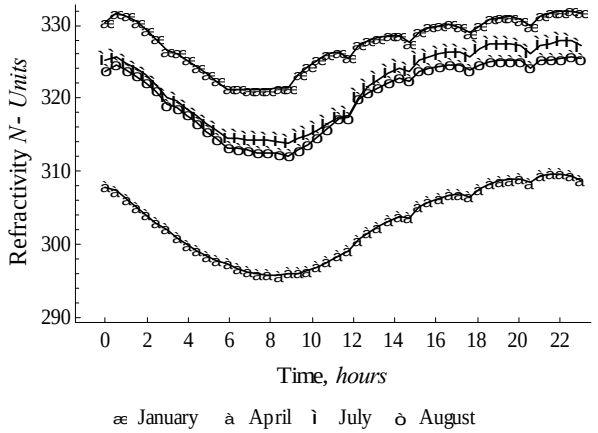


Figure 11. Diurnal variation of radio refractivity in ulaanbaatar

5. Conclusion

The above discussion has emphasized the following points:

a. Radio refractivity shows a seasonal variation with high value in the wet season and low value in the dry season.

b. From the diurnal and seasonal radio refractivity values, the values of refractivity at drier and warmer localities are lower than colder and wettest region.
c. Radio refractivity value over Mongolia increases from about 295 N-units gobian zone to about 328 N-units in the north.
d. The yearly maximum radio refractivity, 328 N-units was in Khatgal in the January and minimum one, 295 N-units was in Dalanzadgad in the May.
e. The variation of refractivity from northern Mongolia to gobian zone have a maximum of about 33 N-units.
f. Diurnal variation of radio refractivity shows that the values of N the highest at midnight and gradually decrease until sunrise and gradually increase after about 09.00 hours.

6. References

[1] Bean, B.R "The Radio Refractive Index of Air", *Proc*, I.R.E., 50, March 1962. pp.260-273.

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

[3] ITU-R, "The radio refractive index: Its formula and refractivity data", 2003, pp.453-459.

[4] Bean, B. and B.A Cahoon "Correlation of monthly median transmission loss and refractive index profile characterestics", *J.Res. N.B.S.*, Vol.65, No.1, 1961, pp.67-74.

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

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

Constructing a system for monitoring, managing groundwater in the industrial zones of Hanoi city

Quang Hiep Vu[1], Thi Hong Nhan Vu[2], Keun Ho Ryu[3]
[1]*Institute of Geophysics, VAST A8-18 Hoang Quoc Viet,Caugiay Hanoi, Vietnam*
hiep88@dblab.chungbuk.ac.kr
[2]*Human Machine Interaction Laboratory UET, Vietnam National University Hanoi, Vietnam*
vthnhan@gmail.com
[3] *Database/Bioinformatics Laboratory, Department of Computer Science,*
Chungbuk National University, South Korea
khryu @dblab.chungbuk.ac.kr

Abstract

Ground water plays an important role in daily activities and production of human. It is the major used water source for Vietnam and industrial purpose. We build a system to collect, and analyze the spatiotemporal data of hydrological and hydro-geological characteristics of shallow and deep groundwater aquifers in northern Hanoi industrial zones and in nearby Red River water. Groundwater level, electrical conductivity were measured in four monitoring wells, complemented by anion, cation, and stable isotope analyses of ground and surface water. The system also could help us: understand how aquifer systems work, determine direction and gradient of groundwater flow, determine annual and long-term changes of groundwater in storage, estimate recharge rate, gain insight for well construction and where to set pump bowls for efficient extraction, as well as predict the water level, water quality in future.

1. Introduction

Because of urbanization, industrialization, and agricultural development rapidly in Vietnam, surface water in lakes, rivers, are increasingly polluted which in turn negatively impacts the quality of groundwater [1]. Groundwater resources are important in supplying water for domestic and industrial use. Especially, over large regions such as capital city of Hanoi. These resources are limited, so it is important to introduce a plan for efficient management by studying spatial and temporal variations of aquifer quality.

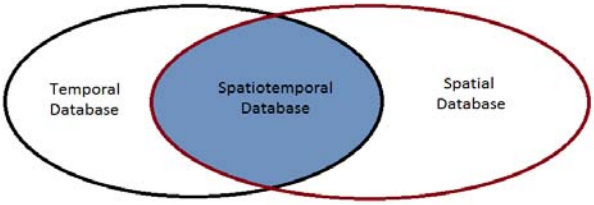


Figure 1. Relationship between spatial, temporal and spatiotemporal DB

Spatiotemporal databases aim to provide database support for applications which show both spatial and temporal characteristics. It provides extensions to existing models of Spatial databases to include temporal aspects, in order to better cater to dynamic environment, like moving objects, traffic flow etc. Spatiotemporal databases can be defined as a database that embodies spatial, temporal and spatiotemporal database concepts and captures both spatial and temporal aspects of data. In real world applications often time and space exists together, hence dealing with spatial aspect without considering temporal aspect is of limited use. Figure 1: provides an idea about the relationship, between spatial, temporal and spatiotemporal database.

This paper introduces a system for monitoring and managing the conductivity and groundwater level varies continuous-time in the Northern Hanoi industrial zones, Vietnam. Then the paper study a model for spatiotemporal database management system. The diagram definition of spatiotemporal database also will be described into details[2]. The method for analyzing the topology relationships between binary geographical objects, especially the

model: 9-intersection [2] will be presented for calculating the relationship of objects. For providing an interaction help with users efficiently and quickly, this paper designs operators related to search conditions of spatial, space and spatiotemporal data. Several illustration examples for that operators will be implemented. A system is finally built with that operators was embedded into the system. The experimental results can then be utilized as a convenient efficient system for monitoring, managing groundwater in the Northern Hanoi industrial zones, Vietnam.

The rest of the paper is organized as follows. Section 2 overviews related work related. Section 3 explains to design the database system. Section 4 will apply operators for data query . Section 5 shows the experimental results. Conclusion and future work is presented in Section 6.

2. Related Work

2.1 Spatiotemporal database concepts

Spatiotemporal concepts, involves both spatial and temporal concepts. Spatiotemporal database concepts need to supply both spatial and temporal database concepts while implementing spatiotemporal database. Hence it is essential to understand the different database concepts needed to create spatiotemporal database systems, they are:

- Spatial databases
- Temporal databases
- Spatiotemporal databases

Spatial databases are designed to store and process spatial information systems efficiently. Spatial databases store the spatial attributes, which have space related properties usually the temporal extent is not present. The fundamental spatial data types are **point, line and region** shown in figure 2 as well as their relationships like **intersects, lies within, lies outside, touches** shown in figure 3.

Figure 2. Fundamental data types in spatial DBMS, point line and region

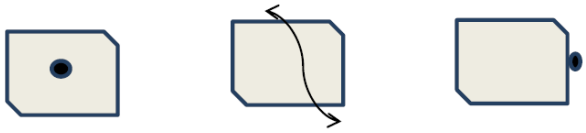


Figure 3. Few relationships between spatial, objects covered, intersects and adjacent.

Temporal databases represent attribute of objects which are changing with respect to time i.e. like functions with continuous range or functions which represent discrete values at different points in time.

2.2 Spatiotemporal groundwater database models

The fundamental aspect of spatiotemporal groundwater systems, is the spatiotemporal groundwater data models. These define entity data types, relationships, operations and rules to maintain database integrity for spatitemporal entities. It must also provide efficient support for spatiotemporal queries and analytical methods to be performed in the spatiotemporal information systems. When designing temporal data models, center, temporal operations, time density and time representation need to be taken into account , whereas for spatial data models, orientation, direction, structure of space,etc need to be considered.

Different models have been proposed for modeling spatio-temporal information systems, such as: The Snapshot Model, Space-Time composite, Simple Time Stamping, Event-Oriented model, three domain model, History Graph Model, Spatio-Temporal Entity-Relationship, Object-Relationship (O-R) Model, Spatio-temporal object-oriented data models, Moving Object Data Model.

Recently, there have been several researches on groundwater management [3, 4, 5, 6]. These works have some weaknesses in the following:

These works almost focus to analyze the statically data collected. The works didn't have visualization model to query the data changes over time and also didn't predict the groundwater level, groundwater quality in future. Hence, this paper will build a spatiotemporal database model for monitoring and managing the groundwater data, the data collected are then used for data visualization. The paper also build operators to query data changes over space and time. Finally, a experimental system for managing the wells in the industrial zones of Hanoi city will be deployed.

3. Design the database system

3.1 The structure of system

In this paper, we design an architectural system to monitor the groundwater information for using the data efficiently.

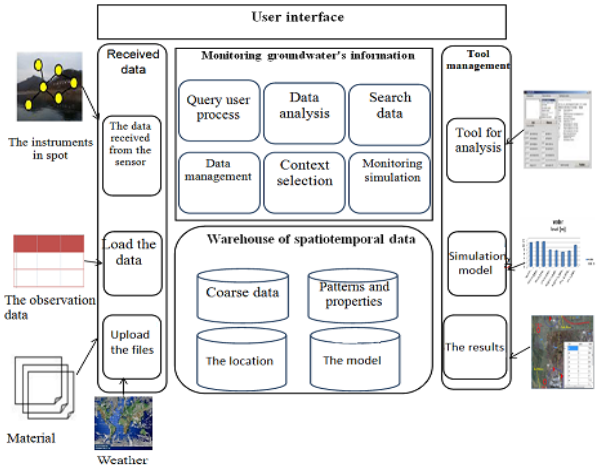


Figure 4. The system to monitor and manage the ground water

The architecture includes many modules as shown in figure 4 such as: location monitor module, collection module, data preprocessing module, spatiotemporal data warehouse module, module for monitoring the groundwater's information, module manages the tool for analysis and information visualization with user interface.

3.2 Description of system

Effective groundwater management will protect the quantity of groundwater and ensure a dependable and affordable supply of groundwater into perpetuity. It will also protect the water quality to ensure that the groundwater remains suitable for domestic, industrial, agricultural, and environmental uses and it will seek to prevent land subsidence that can damage expensive public and private infrastructure such as water conveyance and flood control facilities, and water wells.

Each well is determined by a point and assigned by a label. In this paper, We use the sensors to measure the groundwater level and groundwater quality of each well. It will help us to understand how aquifer systems work, determine direction and gradient of groundwater flow, determine annual and long-term changes of groundwater in storage, estimate recharge rate, gain insight for well construction and where to set pump bowls for efficient extraction, as well as predict the water level, water quality in future.

3.3 Design database

As the data described above, this paper designs the entity relationship (ER) as shown in figure 5:

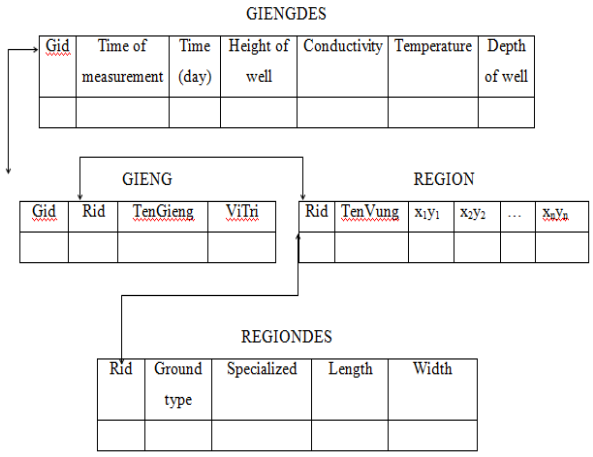


Figure 5. Entity relationship (ER) model

GIENG entity includes the attributes that: Gid, Rid, TenGieng and Vitri. Gid is the primary key of GIENG entity to distinguish with other well, Rid is the foreign key references to REGION entity. TenGieng is the name of well, Vitri is the position of well.

The second entity that GIENGDES describes specifically GIENG entity, this entity includes the attributes: Gid is the foreign key references to GIENG entity, time of measurement, time, height of well, conductivity, temperature, depth of well

The attributes of REGION entity are: Rid is the primary key of REGION entity, TenVung is the name of region, the border of region is a set of points that:x_1,y_1; x_2,y_2,... x_n, y_n

The attributes of REGIONDES entity are: Rid is the foreign key to references to REGION entity, specialized, length and width.

4. Apply operators for data query

In spatio-temporal domain, several operators for logical relationships, arithmetics, position, orientation, extent, surface area, volume, shape and perimeter, spatial topology and set oriented operators like union, cardinality, intersection etc need to be studied. In the case of spatio-temporal operators, in addition to the above mentioned spatial operations, operators related to temporal operations like, event duration,

intersection, union, negation and comparison need to be taken into account.

This paper use the operators have built and embedded in the system such as: time operator for two region, spatial operator for point and region and spatiotemporal operator.

Function: Boolean ST_Operators(R, S/P, Top, Sop)
Input: 1st region (R), 2rd region (S)/point (P), time operator (Top), Space operator (Sop)
Output: st_predicate: boolean
Begin
st_predicate=FALSE;
Step 1: Check time condition
- Select time gap between 1st region and 2rd region (VTs, VTe), (begin, end)
- Call function Temporal_Operators() with parameters Top, VTs, VTe, begin, end;
Step 2: check space condition
If (Temporal_Operators(VTs, VTe, begin, end) = 'true' **then**
If space condition is region-region **then**
- call function S_Operators() with parameters R, S, Sop;
- **If** (S_Operators()=TRUE **then** *st_predicate* = TRUE
If space condition point-region **then**
- call function isContained() with parameters P, R;
- **If** (isContained()=TRUE **then** *st_predicate* = TRUE
endif
Return *st_predicate*
End

Algorithm 1. Spatiotemporal operator

Hence, the user easily find the data without remember the operators as shown in the Table I:

Table 1. Comparison the algorithm have operators and without operators

(a) Without operators	(b) Have operators
Select GIENGDES.Gid, GIENGDES.DoSau From GIENG, GIENGDES, REGION Where (GIENGDES.Gid=GIENG.Gid) and (GIENG.Rid=REGION.Rid) and (REGION.TenVung = Quang Minh) and (QuangMinh.x1<= GIENG.x) and (GIENG.x < QuangMinh.x2) and (QuangMinh.y2 <= GIENG.y) and (GIENG.y < QuangMinh.y1) and (W.x1<= GIENG.x) and (GIENG.x < W.x2) and (W.y2 <= GIENG.y) and (GIENG.y < W.y1) and (V.VTe < 1/1/2009)	Select GIENGDES.Gid, GIENGDES.DoSau From GIENG,GIENGDES, REGION Where (GIENGDES.Gid=GIENG.Gid)and (GIENG.Rid=REGION.Rid) and (REGION.TenVung = Quang Minh) andST_Operators(REGION.QuangMinh,W,overlap,(VTs,VTe) before (1/1/2009,now))

5. Experiment and results

This section presents the experimental settings as well as shows several experimental results.

5.1 The experiment integrates time operators, space operators and spatiotemporal operators

This paper built a system with the experiments integrate time operators, space operators and spatiotemporal operators.

Figure 6. The user interface

The role of operators are the filter to find entity. The priority of data query follow order: time, space. That mean when the user finds the entities have both these conditions, the entities satisfy the time condition then will be continued to check the space condition. Hence the time for finding the entity will be reduced because to find the entity in the spatiotemporal database is very complicated and need a lot of time.

There are several experimental examples using the operators for finding the entity.

Figure 6 shows the user interface uses to input the values

The user interface have the input values:

- **Khu Cong Nghiep:** use to choose the Region
- **TT khong gian:** the type of space operator: **IsContained.**
- **TG bat dau t1** and **TG ket thuc t1** are: start and end of the interval time T1 of time operator's name: **during**
- **TG bat dau t2** and **TG ket thuc t2** are: start and end of the interval time T2 of time operator's name: **during**
- **TT Thoi gian:** the type of time operators**: overlab**s and **during**

5.1.1 Data query with space conditions

Figure 7. Input values with spaces condition

Input values are shown in figure 7:

- **Khu Cong nghiep:** Bac Thang Long
- **TT khong gian:** Iscontained to find the position of wells in the table: GIENG in Bac Thang Long area have MBR (Rx1, Ry1, Rx2, Ry2)

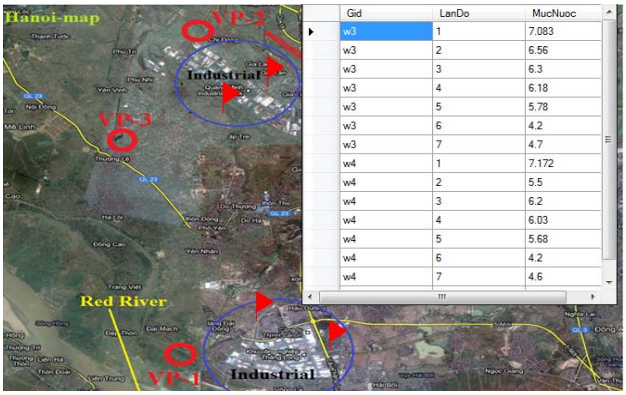


Figure 8. Output values with space condition

Output values are shown in figure 8 include the groundwater level of two wells in the Bac Thang Long area: w3 and w4.

5.1.2 Data query with time condition

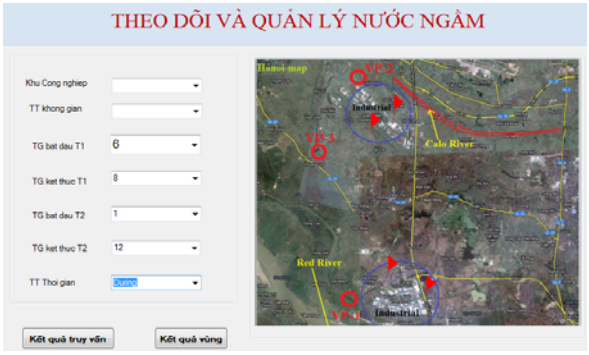


Figure 9. Input values with time condition

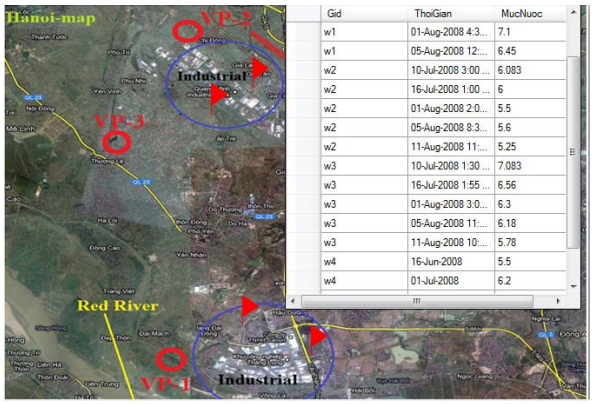


Figure 10. Output values with time condition

Figure 10 shown the results to query groundwater level in two intervals time: T1='6/2008, 8/2008' and T2='1/2008,12/2008' use time operator: **during.** With this query condition, the groundwater level of all wells (four wells: w1, w2, w3, w4) are shown.

5.1.3 Data query with spatiotemporal condition

Input values are shown in figure 11

Figure 11. Input values with spatiotemporal condition

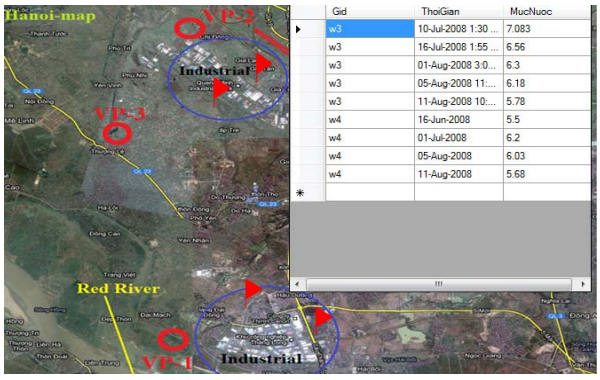


Figure 12. Output values with spatiotemporal condition

Figure 12 shown the results to query groundwater level of two well (w3, w4) in two intervals time: T1='6/2008, 8/2008' and T2='1/2008,12/2008' use space operator: **IsContained** and time operator: **during.** With this query condition, the groundwater level of all wells (w1, w2, w3, w4) are shown and satisfy the time condtion

6. Conclusion and future work

The WHO (World Health Organization) have warned that lack of open access to clean drinking water is the one of the major adverse influences on the general health and life expectancy of people in many developing countries. With the development of information technology has been a main factor for determining groundwater information. A lot of works have reported that urbanization, industrial and agricultural activities directly or indirectly affect groundwater. However, there is still very little information on this quality in Vietnam and this should be redressed.

Spatiotemporal databases have become very important in recent years, as many real world applications, one of important application is hydrogeology information. The hydro-geological data have attributes related to both space and time, and managing them using existing database management system is complex and inefficient, as these entities which show spatiotemporal behavior are multidimensional. Hence, this paper built a system use the spatiotemporal operators embedded in the system to query the groundwater data in northern Hanoi industrial zones. The system is very convenient and efficient for knowledgeable about aquifer systems work, direction and gradient of groundwater flow, annual and long-term changes of groundwater in storage, recharge rate, etc, as well as help us make a plan to manage groundwater efficiently.

In ongoing work, I will build a system for predicting the groundwater level and groundwater quality in future, it will help us to warn the status of groundwater resource better.

7. References

[1] Dzung NT, "*Current status of groundwater pollution in Hanoi area. Proceedings of the International Symposium on Environment and Injure for Community Health Caused by Pollution during the Urbanization and Industrialization*", Hanoi Dec. 28–29, 2002, pp. 55–69

[2] M.J.Egenhofer, R.Franzosa, "*Point –set Topological Spatial Relations*", Int J GIS, Vol.3, 1991, pp. 161-174.

[3] P.S.R.Kiran, R.T.Kumar, K.Stanlin, P.Archana, L.Sridevi và A.S.Radha, "*GIS Techniques for Groundwater contamination Risk Mapping*", Volume 20, Issue 2-3, November 1999, pp. 279-294

[4] W.T.Fang, K.M.Lin, C.V.Chin và H.Y.Lu, "*Geographic information system applied in the groundwater*", Advancing Water Resources Research and Management, September, 1996.

[5]Munda A, Haiduk A, "*Hydrochemical characteristics of groundwater in the Kingston basin, Kingston, Jamaica*", EnvironEarth, 2011, pp. 415–424.

[6] Bui DD, Kawamura A, Tong TN, Amaguchi H, Nakagawa N, "*Spatio-temporal analysis of recent groundwater-level trends in the Red River Delta, Vietnam*", Hydrogeol, 2012.

Investigation of SEE on a 32-bit microprocessor based on SPARC V8 architecture by laser test

Chunqing Yu, Long Fan, Suge Yue, Maoxin Chen, Shougang Du
Beijing Microelectronics Technology Institute
yuchunqing2006@163.com

Abstract

In this paper we use a picoseconds pulsed laser facility to investigate the SEE (single event effects) of CVLSIC (Complex Very Large Scale Integrate Circuits). A 32-bit microprocessor based on SPARC V8 architecture has been chosen to explore the contribution of different parts to the single event effect in details. The sensitivity mapping of cache, regfile and the sensitivity thickness of devices which are fabricated in 180nm technology, with a substrate thickness of 400μm are obtained. In this way, three-dimensional distribution of the sensitive areas is obtained which offers essential information to the radiation hardening work in the future.

1. Introduction

Pulsed laser technology is a new experimental verification technology developed in recent years, it offers numerous advantages for SEE testing compared to conventional testing. For example, laser testing doesn't cause degradation, has lower cost and is repeatable, etc. Nowadays the use of a pulsed laser beam as a test and hardness assurance tool to simulate single event phenomena (SEP), such as upsets (SEU) or latch-up (SEL), in microelectronic devices has been becoming more and more popular [1]. The laser facility can assess the hardened level and may ensure a faster detection and diagnosis of radiation sensitivity weaknesses. The laser pulse can be used to characterize sites on the same chip having specific sensitivities to radiation impact. Moreover, the laser pulse can be synchronized with circuit clocks and, thus, can be used to characterize the influence of dynamic circuit operation on its sensitivity to radiation.

Currently, a portion of devices have been conducted laser experimental for single event effects study ,including SRAM, FPGA, optoelectronic devices, etc. Whereas studies of the sensitive thickness of sensitive areas for CVLSIC by pulsed laser at home and abroad are rarely seen in published literature. Therefore, in this paper a detailed study on the performance of the SEE of microprocessor based on picoseconds pulsed laser equipment were carried out to analyze the discrepancy between different functional units of the device. Simultaneity the curves of sensitive thickness in different sensitive sites of various functional units are presented. Accordingly pulsed laser test as an effective and economic experiment method will be used to estimate the reliability and capability of domestic CVLSIC in the near future.

2. Background

Experimented circuit is a 32-bit embedded RISC microprocessor based on SPARC V8 architecture illustrated in Figure 1 which can be used for on-board embedded real-time computer system, to meet a variety of aerospace missions. We just need to add memories and related peripheral circuits to form a complete single board computer system. The target circuit was fabricated in bulk silicon CMOS 180nm technology, with a substrate thickness of 400μm.

The operating frequency of microprocessor is 100MHz. Device has many different function blocks among which I-cache, D-cache, Regfile, IU/FPU units are susceptible to upsets [2]. It has six metal layers, inability to penetrate metal overlays, making the sensitive areas beneath them difficult to access. Therefore a new approach has been developed to interrogate SEE phenomena through the wafer using backside irradiation, eliminating any interference from the metallization layer stacks [3]. Laser test system is illustrated in Figure 2.

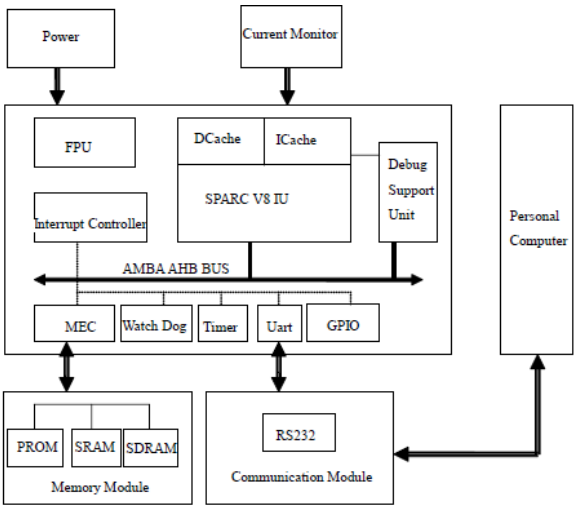


Figure 1. Architecture of the microprocessor

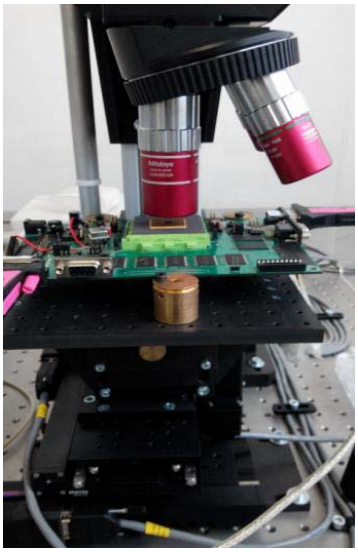


Figure 2. Laser test system

3. Experiment details

A picoseconds pulsed laser system is used to perform the tests. It generates pulses of 1064nm wavelength and 17 ps temporal optical duration, operating either in a single-shot mode or with a high repetition rate of 1k Hz. The chip is connected to a test board mounted on a motorized xyz stage with 0.125μm resolution. The optical pulses are focused onto the backside of the target circuit with a 100× microscope objective. The diameter of the focused laser spot is approximately 1.5 microns at the sample surface.

Several areas selected for scanning are illustrated in Figure 3. When tested, corresponding test programs are implemented to explore the sensitivity in different function units. During scanning the selected area, any errors in the contents of the memory units are recorded.

Large step increases in the pulse energies are used, to ensure quick detection and location of SEP. A subsequent SEP characterization algorithm is then performed for each detected sensitive site. This consists in determining the laser pulse energy threshold that triggers the event, with the pulsed laser beam in the single-shot mode [4]. This measurement has been performed for different locations in the sensitive site area with a certain value of Z-Axis offset which is the distance of laser focal plane from it reaches the surface of the substrate to the deep inside the device.

As the laser focus plane close to the sensitive volume, laser energy required for triggering SEE is reducing and when the laser focal plane move away from sensitive volume, laser energy required for triggering SEE is gradually increasing. So during the section that laser focal plane just reaches the sensitive volume to it departs from the sensitive volume, the energy required for triggering an SEE is the minimum. Record the sensitive sites, laser threshold energy and then plot them in a curve. Then we change the value of Z-Axis offset. The above tests are repeated for a range of different values of Z-Axis offset sufficient to define the optimal Z-Axis value when the trigger energy is lowest. In this way, we get the E-Z curves in different sensitive sites.

A 20μm×20μm cache/regfile area is selected for scanning. The scanning step is 1μm per second. This test is done under various laser energies. Record the coordinate where occurs SEE and then plot them in a curve, we obtained the sensitive mapping of different units.

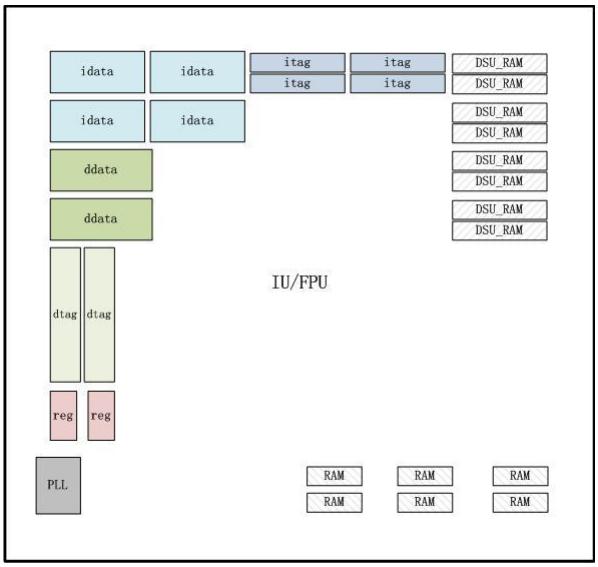


Figure 3. Scan areas sketch map

4. Results and Discussion

4.1. E-Z curves in different sensitive sites Change the focus depth namely Z-Axis offset

we can obtain the sensitive nature of the DUT. We have obtained the SEE data of several sensitive points in cache areas, regfile areas and IU/FPU areas.

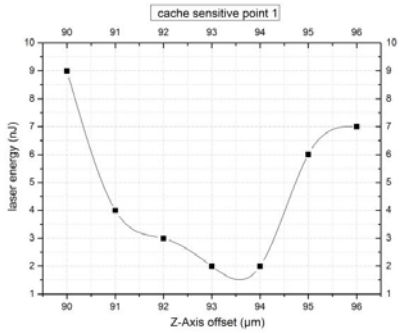


Figure 4. The sensitive thickness of sensitivity site 1 in cache area

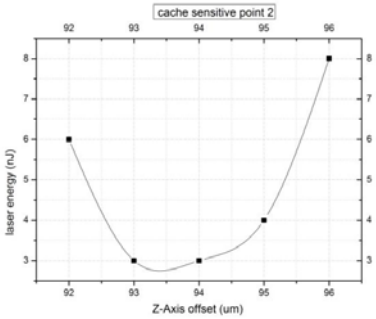


Figure 5. The sensitive thickness of sensitivity site 2 in cache areas

As is shown in Figure 4 and Figure 5, the horizontal axis is the distance of laser focal plane from it reaches the surface of the substrate to the deep inside the device and the vertical axis is the threshold laser energy to trigger an event. During the section that laser focal plane just reaches the sensitive volume to it departs from the sensitive volume, the energy required for triggering an SEE is the minimum. So the distance of two dots which have the lowest laser energy is the thickness of sensitive area. From the above figures, we can see that the thickness of cache sensitive area is about 1μm.

Figure 6 and Figure 7 is the SEE data of regfile area, in which the horizontal axis and the vertical axis have the same meaning as in Figure 4 and Figure 5. As can be seen from the Figure 6 and Figure 7, the sensitive thickness in regfile area is about 1μm.

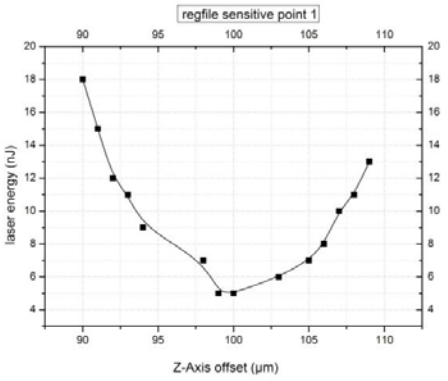


Figure 6. The sensitive thickness of sensitivity site 1 in regfile area

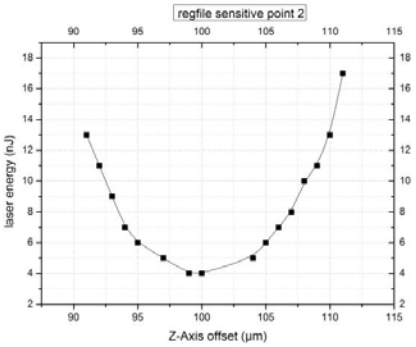


Figure 7. The sensitive thickness of sensitivity site 2 in regfile area

Similarly to the above, Figure 8 is the SEE data of IU/FPU area, in which the horizontal axis and the vertical axis have the same meaning as in Figure 4 and Figure 5. As can be seen from Figure 8, the sensitive thickness in IU/FPU area is about 1μm.

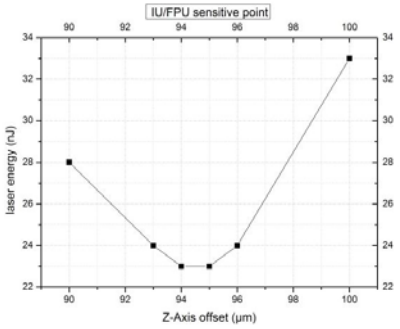


Figure 8. The sensitive thickness of sensitivity site in IU/FPU area

In short, as can be seen from the above results, we can draw a conclusion that the sensitivity thickness of device is about 1μm which is fabricated in 180nm technology, with a substrate thickness of 400μm.

4.2. Sensitive mapping of different function units

The scatters of test results under different energies are illustrated in Figure 9. It implies that the number of errors increases with the increase of laser energy. Then the test results under different conditions are plotted in a curve. In this way, we obtain the sensitive mapping of cache which is shown in Figure 10.

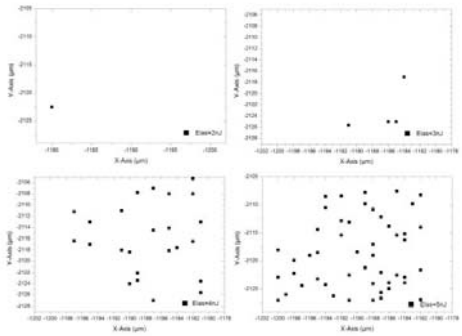


Figure 9. Scatters of cache

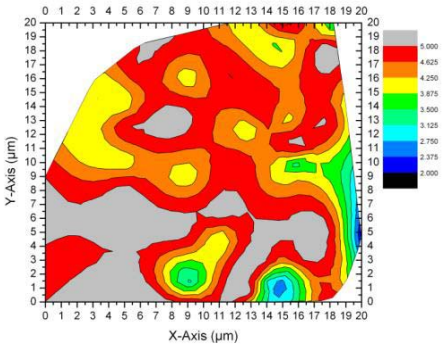


Figure 10. Sensitive mapping of cache

In Figure 10, the X-Axis and Y-Axis is the coordinate of horizontal axis and the vertical axis that occur SEE. Different colors represent the probability of SEE. As can be seen from Figure 10, black color area has the greatest probability of SEE and the gray color area has the relatively smaller probability of SEE. The laser threshold energy of cache is about 2nJ.

Similarly, we obtained the scatters of test results under different energies which are illustrated in Figure 11 and the sensitive mapping of regfile shown in Figure 12. In Figure 12 the X-Axis and Y-Axis and the different colors have the same meaning as in Figure 10. As can be seen from Figure 12, the laser threshold energy of regfile is about 7nJ.

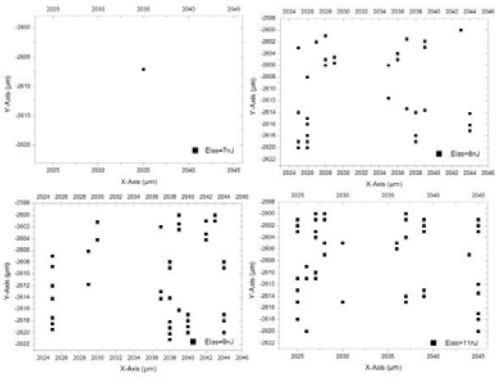


Figure 11. Scatters of regfile

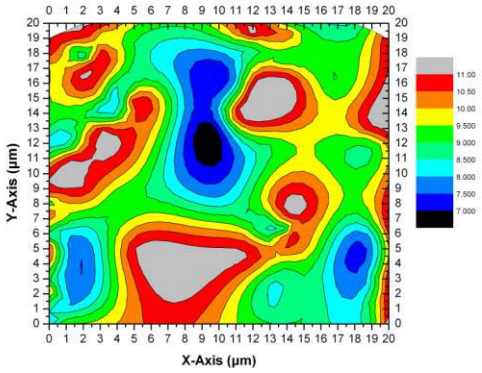


Figure 12. Sensitive mapping of regfile

The test results imply that cache has lower laser threshold energy than regfile. The difference of sensitive mapping between cache and regfile may be caused by difference of layout. Simultaneity two-bit upsets and multi-bit upsets are detected which may be induced from charge sharing. This phenomenon becomes more apparent along with the increasing laser energy.

4.3. Discussion

The reasons for the deviation of individual experiment points in the curves can be mechanical vibration error, human operator error, positioning error, or the flatness error of the device substrate, etc. However, no matter how much the optimal Z-Axis offset is, the thickness of the device sensitive areas is about 1μm. Synthesizing the sensitive mapping and the thickness of sensitive areas, we obtain the three-dimensional distribution of the sensitive areas which offer essential information to the radiation hardening work in the future.

5. Conclusion and future work

The use of a focused picoseconds pulsed laser system can generate SEE phenomena rapidly and efficiently. We can draw a conclusion from the laser energy and Z-Axis offset curves that devices fabricated in 180nm technology, with a substrate thickness of 400μm, have a 1μm thickness of sensitive volume. Also we get sensitive maps of different units. Finally, three-dimensional distribution of the sensitive areas is obtained which offers essential information to the radiation hardening work in the future. Summing-up, the approach of using an array of pulses covering a selected die area provides a new solution in investigating the reliability and capability of domestic CVLSIC before they are applied in spaceflight [7].

The next step is to perform a different test vectors under various duty factors to obtain the relationship between duty factor and the SEU sensitivity. So we can predict a device's SEE sensitivity through analyzing its programs which will be used in space application.

Acknowledge

This work was supported by the following Projects: test evaluation methods research for anti-radiation performance of nanometer ICs and the SEE verification and assessment techniques of CVLSIC based on laser micro beam.

6. References

[1] Zhifeng Lei H L H C. "Single Event Effects test for CMOS devices using 1064nm plused laser[J]" *IEEE 978-1-4577-1232-6*, 2011.

[2] J.H.Elder J O W A, "A method for characterizing a microprocessor's vunlnerability to SEU [J]", *IEEE Transaction on nuclear science*, Vol.35, No.6, December 1988.

[3] Darracq F L H B N. "Backside SEU laser testing for commercial off-the-shelf SRAMs [J]", *IEEE Trans Nucl Sci*, Vol.49, No.6, 2002, pp.2977-2985.

[4] James R.Schwank E. "Estimation of Heavy Ion LET Threshold in Advanced SOI IC Technologies from Two-Photon Absorption Laser Measurements [J]," *IEEE Trans.Nucl.Sci*, Vol.57, No.4, August 2010, pp. 1827-1834.

[5] R. Koga, W. A. Kolasinski, and M. T. Marra Space Sciences Laboratory, The Aerospace CorporationP. 0. Box

92975, Los Angeles, CA 90009W. A. Hanna McDonnell Douglas Astronautics CompanyP. 0. Box 516, St. Lous, MO 63166 “Technology of microprocessor testing and SEU rate prediction”, *IEEE Transactions on Nuclear Science*, Vol. NS-32, No. 6, December 1985.

[6] R. Velazco , T. Calin , M.Nicolaidis. S. C. Moss, S. D. LaLumondiere, V.T. Tran4, R. Koga, “SEU-Hardened Storage Cell Validation using A Pulsed Laser,” *IEEE Transactions on nuclear science*, Vol.43, No.6, December 1996.

[7] S. Buchner, D. McMorrow, J. Melinger, and A. B. Campbell, “Laboratory Tests for Single Event Effects”, *IEEE Transaction on nuclear science*, Vol .43, No.2, April 1996.

Session : Interactive Session 3

- Integrated Public Bike Rental System Design
  Sunny Song, Kyu Ik Kim, Myung-Sic Kim, Keun Ho Ryu
- Study on the Method of Composing Data Warehouse for Error Verification and Multi-dimensional Error Test
  Myung-Sic Kim, Gwi-Seop Song, Kenu Ho Ryu
- Wireless Uroflowmetry System for Self-test at Home
  In-Kwang Lee, A-Rong Heo, Ho Sun Shon, Keun Ho Ryu, Kyoung-Ok Kim, Eun-Jong Cha, Kyung-Ah Ki m
- Context Ontology based Mobile Information Retrieval
  Mi Sug Gu, Ho Sun Shon, Keun Ho Ryu
- Protected Health Information for Research on Computer Forensics
  Yoonhwan Shin, Keun Ho Ryu
- Evaluating the Impact of Design Patterns on Code Design using Object-Oriented Metrics
  Batnyam Battulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar
- 3D Reconstruction from Uncalibrated Images
  Tsetsegjargal Erdenebaatar, Suvdaa Batsuuri
- Polynomial Approximation of Impedance of Microstrip Patch Antenna
  Batpurev Mongol, Gerelmaa Byambatsogt, Ganbat Baasantseren

Session : Interactive Session 3

- Hand Gesture Controlled Drawing Tool using "Asus xtion pro"
  Amartuvshin Renchin-Ochir, Dorjnamjirmaa Badraa
- Analyzes of Enrollment Database of a University Information System
  Bulganchimeg.B, Naranchimeg.B, Oyun-Erdene.N, Yanjindulam D, Sodbileg.Sh
- Improvement of the Database Performance of a University Information System
  Munkhtuya.D, Naranchimeg.B, Oyun-Erdene.N, Sodbileg.Sh
- Quadrupeds Motion Data Collection Method
  Javkhlan Rentsendorj, Erdenebat Budsuren, Baatarbileg Altangerel, Oyun-erdene Namsrai
- Installment For Measuring And Sharing Pm 2.5 Air Polution Concentration Through Social Media
  Unursaikhan Batbayar, Sereeter Lodoysamba, Christa Hasenkopf, Joe Flasher

Integrated Public Bike Rental System Design

Sunny Song, Kyu Ik Kim, Myung-Sic Kim, Keun Ho Ryu
Database/Bioinformatics Laboratory, Department of computer science,
Chungbuk National University, South Korea
thdtjsgmfr@naver.com, ora011@nate.com, khryu@dblab.chungbuk.ac.kr

Abstract

The PBRS(Public Bike Rental System) is a bike sharing system operating to rent a bike rental station installed at key points in the city and after moving to the destination, then to return to the nearest bike station after moving to the destination. That now contribute to public transport activation and individual ′s physical strength. But the current system integrated management is difficult because each operated by a municipality and deteriorate the quality of service. Therefore in this paper designed Integrated Public Bike Rental System as national standard.

1. Introduction

The PBRS(Public Bike Rental System) achieved a nation′s vision that low-carbon and green growth is a bike sharing system operating to rent a bike rental station installed at key points in the city and after moving to the destination, then to return to the nearest bike station after moving to the destination.

PBRS started in the way of environmental protection but now contribute to public transport activation and individual′s physical strength. Also recently, the Clean Development Mechanism is a global concern and the trend is enforce a act of bicycle use and regulations of each municipality so national interest and the continued operation is necessary[1]. But the current system integrated management is difficult because each operated by a municipality and deteriorate the quality of service. Therefore in this paper designed Integrated Public Bike Rental System as national standard.

This paper is structured as follows: In section 2 we mention the current state and problems of PBRS in Korea. Section 3 describe our system design and finally, in section 4 we present our conclusions and future works.

2. Current State and Problems of Public Bike Rental System in South Korea

Local Governments of the South Korea have been adopted public bike rental system for aspects of benefits such as reducing an emission of carbon dioxide with using public transportation, and realizing the promotion of national health using bicycles. Classic examples of Public Bike Rental System in the South Korea are as follow: "Tashu" of Daejeon Metropolitan city, "Nubija" of ChangWon city, "FIFTEEN" of Goyang city, and so on.

According to the analysis of the trend about the rental bikes after adopting Public Bike Rental Service, it has many potential strengths and demands. However this service need to improve the quality[2]. The earlier services have not been supporting a rental of the other local bike between neighboring districts. Particularly this service have been operating 4 areas in Seoul: Sangam, Yeouido, Seodaemun-gu, and Seocho-gu. The service of Sangam and Yeouido have managed at the Seoul City Hall. And others are handled by their borough offices. Figure 1 is shown the state of bike stations in Seoul. But here is marked just Sangam and Yeouido on the map.

Like this Figure 1 clearly show that separated operating bike rental system in one Seoul so they are not interchangeable and have constraints to be used.

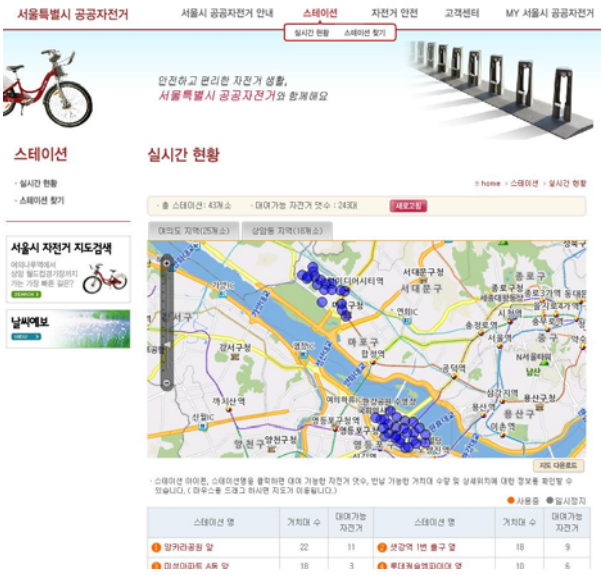


Figure 1. Status of public bike stations in Seoul

Furthermore municipalities operating PBRS have difficulties of maintenance costs such as labor and shipping costs. Meanwhile many countries of the world are trying to reduce greenhouse gases in the atmosphere such as the Convention on Climate Change and Kyoto Protocol, and South Korea also intend to pursue the voluntary reduction to 30% by 2020. Therefore it need to operate to regulate carbon emissions and support measures at national level.

In this paper designed Integrated Public Bike Rental System as national standard for resolve the problem of service, cost and environmental aspects of current separate system.

3. Design of Integrated Public Bike Rental System

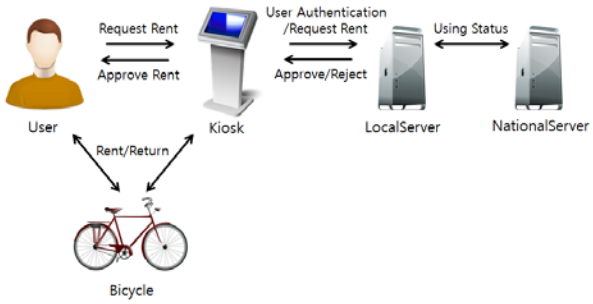


Figure 2. Concept of Public Bike Rental System

Figure 2 present a concept of the proposed system. A user can rent a bike using a Kiosk. And then the rental data is handled at the Local Server. When the system need the national data, the National Server send a request data to Local Server.

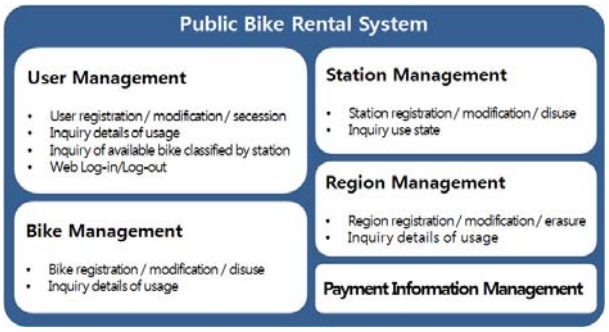


Figure 3. Block Diagram of Public Bike Rental System

Figure 3 is shown a block diagram of the system. This system consist of five parts such as User Management Module, Bike Management Module, Station Management Module, Region Management Module, and Payment Information Management Module. Firstly User Management Module have functions as User Registration/Modification/Secession, Inquiry of available bike classified by station, and Web Log-in/out. Secondly Bike Management perform Bikes Registration/Modification/Disuse. Thirdly Station Management Module have Stations Registration/Modification/Disuse function. Fourthly Region Management Module also have function as Region Registration/Modification/Erasure. Finally Payment Information Management Module manage particulars of payment. In addition, All modules are able to inquire details of usage about each information.

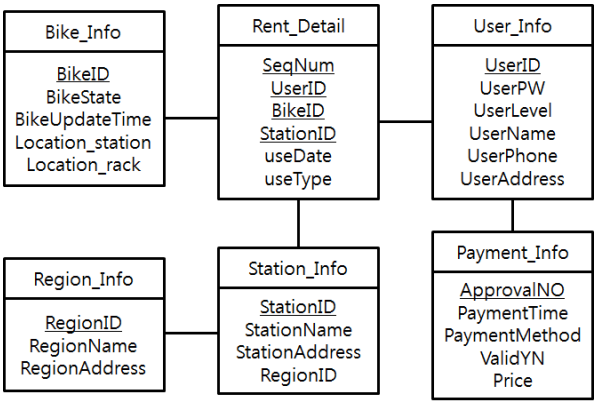


Figure 4. E-R Diagram of Public Bike Rental System

Figure 4 is shown ER Diagram of our system. This system have six tables. The rental history data is stored in Rent_Detail table.

4. Conclusion

We analyzed the status and problems of Public Bike Rental System and were presented and designed integrated public bike rental system to solve this problem. If this designed system adopted the real public bike rental business, standardized systematic management is possible because operating jurisdiction will be changing from local government to the country. Through integrated management system can be achieved to improve the quality of service of membership management, asset management, operations management, accounting systems, etc. And also we expected to increase the share of bicycle transportation.

In the future work, we need to develop the operating system based on the design of this system and then measure actual effect.

5. References

[1] S. Y. Yun, H. S. Kim, S. C. Park, ″Design and Implementation of RFID/GPS-based Public Bicycle Location Management System″, *Proceedings of the 2010 KIIT Summer Conference*, Korean Institute of Information Technology, 2010, pp219-222.

[2] J. Y. Lee, ″The Analysis of Characteristics of Public Bike System and Its Implication on Daejeon City″, Daejeon Development Institute, 2010.

[3] ″The Rental Service of Public Bike in Seoul Metropolitan Government″, ″Seoul Metropolitan Government″, Retrieved from https://www.bikeseoul.com/index.do, 2014.

[4] ″The Public Bike Service in Seocho Seoul Korea″, ″Seocho Seoul Korea″, Retrieved from https://scbike.seocho.go.kr/index.do, 2014.

[5] ″The Unmanned Bike Rental System in Seodaemun-gu″, ″Seodaemun-gu office″, Retrieved from http://bike.sdm.go.kr/status/depository/readDepositoryListAndDetailStatus.xhtml?depositoryId=4, 2014.

[6] ″Bike information of Seoul Transport Operation & Information Service″, ″Seoul Transport Operation & Information Service″, Retrieved from http://topis.seoul.go.kr/renewal/traffic/BikeInfo.jsp, 2014.

[7] E. Gamma, R. Helm, R.Johnson, and J. Vlisside, “Design Patterns: Elements of Reusable Object Oriented Software”, *Addison-Wesley Publishing Company*, Reading, 1995, MA.

Study on the Method of Composing Data Warehouse for Error Verification and Multi-dimensional Error Test

Myung-Sic Kim[1], Gwi-Seop Song[2], Kenu Ho Ryu[3]
[1]*Neoforce Co., Ltd.*
ora011@nate.com
[2]*DataClip Co., Ltd.*
songzsm@naver.com
[3]*Database/Bioinformatics Laboratory, Department of computer science,*
Chungbuk National University, South Korea
khryu@dblab.chungbuk.ac.kr

Abstract

Vast store of database has been established through the project for establishment of database executed by the companies and national institutions. However, there are numerous problems including the difficulties in provision of services to the public and occurrences of civil petitions due to deterioration in the quality of the database. Accordingly, generalized error test method and standardized error verification technology for the database that can be applied under any environment whatsoever are in urgent need. In this Study, the composition of Data Warehouse for multi-dimensional error verification and designing of the multi-dimensional error test system structure are proposed as a means of resolving these issues.

1. Introduction

The error test is performed by using S/W at the data test stage among the process of establishing database. The S/W error test developed cannot be used for error test at other institutions or business sites, and the S/W must be customized for application every time. However, since large scale correction is needed, the costs to be incurred would be similar to the cost of new development. Supplementation through relevant technology development is necessary as there are numerous inadequacies in applying them without corrections. Although vast store of the database has been established through the project for the database establishment executed by the national institutions, problems such as deterioration of the database quality and lack of compatibility of S/W error test arising from the quantitative expansion of thedatabase. Therefore, there is an urgent need to define and establish generalized error testmethod and standardized technology that can be applied in the establishment of the data base. In this Study, the literature on the means of general composition of Data Warehouse and quality factors as well as definition of the data will be examined, and structural design for the multi-dimensional error verification system will be proposed.

2. Literature review

2.1 Date processing method of Data Warehouse

Corporations or public institutions with sales revenues or work activities that are above the prescribed level are those who frequently establish Data Warehouse. Data Warehouse is the database for work analysis composed in the format of processed data by regularly extracting information that users need from the database system that processes the work transactions. Corporations must acquire strategic information and forecast the future by using the information they acquired in order to survive and grow in the midst of rapidly changing corporate environments. For example, in the marketing domain, various attempts are made to achieve the effect of maximizing the sales revenue and elevate the level of satisfaction of the customers by analyzing the mid to long-term consumption trend of the customers, analyzing and assessing the characteristics and preference level for the products of certain customer groups, and analyzing the temporal and regional demands by establishing Data Warehouse. [1]

Important common factor discovered from the experience of establishing the Data Warehouse is the lower quality of the transaction information inputted into the information system during the process of conducting business in comparison to the expected quality. Therefore, the process of Data Warehouse establishment includes the process of cleansing the data. That is, various measures are taken to secure the quality of the data extracted from the work system prior to storing the data in the storage location of the Data Warehouse . These procedures are referred to as the Extraction, Transformation and Loading (ETL) process in the domain of the Data Warehouse illustrated in the Figure. 1.

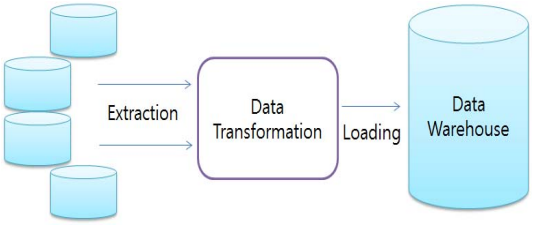


Figure. 1 ETL Process in the Data Warehouse

This is a process of transferring the various data dispersedly stored in multitude of source database to the storage location of the Data Warehouse. It is possible to establish the Date Warehouse by extracting and loading accurate data through the ETL process. [3][4]

2.2 Definition and elements of data quality

With rapid increase in the Informatization of the works of corporations, the problem of severe redundancy and non-concordance of data between the information systems for each of the domains of the works is being highlighted. For example, although an increasing number of organizations are operating the Data Warehouse which integrates the data at the company-wide level in order to effectively support the decision making by the information system users, erroneous data loaded into the information system in the process of integrating and operating the dispersed information system are not being managed properly. Damages are occurring due to wrong decision making arising from such erroneous data and the dissatisfaction of customers resulted from low quality data is also increasing. Data quality is the measure of the consistent satisfaction of the expectation of the users. Here, the term consistency refers to the satisfaction of the expectation of the majority of all users rather than a small portion of users, signifying that the maintenance of concordance between the data is of great importance. It is expected that the data quality will enhance the level of the satisfaction among the users and the interested parties, and, as such, the level of expectation needs to be evaluated by the satisfaction level of the users. The elements of data quality in the Data Warehouse include Accuracy, Completeness, Consistency and Timeliness. [1][2]

- Accuracy: Signifies whether the value contained in the database accurately coincides with the True Value.
- Completeness: Signifies whether the value of the data demanded to satisfy the business demands is being held.
- Consistency: Signifies that the state of consistency is maintained without two or more data being in conflict.
- Timeliness: Signifies whether the data of the Data Warehouse are being updated with the latest data. Service procedures are as follows.

3. Designing of Error Verification System

3.1 Generation of Data Warehouse for error verification

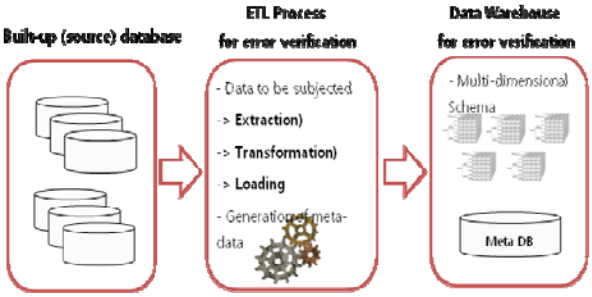


Figure. 2 Flow chart for generation of Data Warehouse for error verification

The flow process of the generation of Data Warehouse for error verification has to undergo the common ETL process of Data Warehouse as illustrated in the Figure. 2. This is the process of generating the Data Warehouse for error verification through extraction, conversion (transformation), integration and loading of the data to be subjected to error verification at the built-up (source) database. Form and items to be subjected to error verification will be the subject of extraction, and the subject data will be extracted by selecting the data from the immediately preceding date or from the past as the subject. Then, the subject items of error verification are selected, enriched, and transformed by separating and integrating the items subjected to verification in accordance with use.The detailed processing

procedures and contents of ETL process are separately registered in the Meta database for management.

3.2 Composition of Data Warehouse for error verification

The composition of the Data Warehouse Module for error verification is illustrated in the Figure. 3.Provision of multi-dimensional perspectives on the data becomes the basic foundation for the efficient error verification and analysis by the user through intuitive and quick understanding of the data with complex relationship.The structure of the multi-dimensional error verification perspective is composed of more than one dimension and item value, and provides multi-dimensional perspective to the user

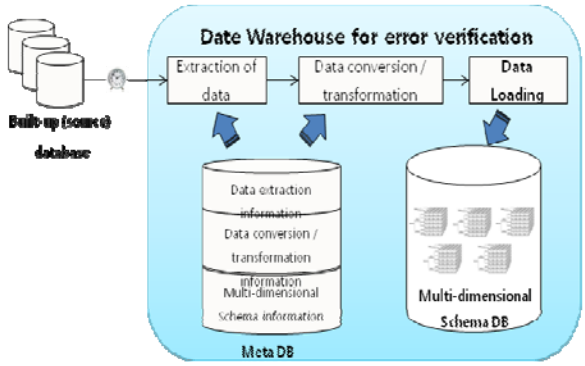


Figure. 3 Composition of Data Warehouse Module for error verification

- Management of linking of built-up (source) database: Function for setting of access information for the built-up (source) database, and manages database IP Address, account information and service ID, etc.
- Extraction of subject data: Select and extract the forms and items that are the subjects of error verification from the built-up (source) database.
- Conversion of subject data: Select and enrich the items to be subjected to error verification, and convert (transform) the items to be subjected to verification subject for the use.
- Loading of subject data: Once the Meta data is generated by extracting and converting the data, load the subject data onto the built-up (source) database.
- Management of Meta data: Manage the data extraction and conversion, and multi-dimensional schema generated information.
- Management of multi-dimensional schema: Manage the multi-dimensional schema composition information of the items to be subjected to error verification.

3.3 Composition of multi-dimensional error verification system

Figure. 4 illustrates the composition of module for the multi-dimensional error verification system where the algorithm that converts the multi-dimensional error verification inquiries into the relationship-type inquiries is applied. The inquiries are the multi-dimensional error verification inquiries composed on the basis of the intuitive understanding from the perspectives of the user and to be used at the time of error verification. However, the data that have to be subjected to the actual error verification is stored in the Data Warehouse for error verification in the format of relationship-type database. In order to obtain the outcome of multi-dimensional error verification inquiries from the relationship-type database, the module and algorithm that convert the relationship type inquiries into SQL are needed. In addition, in order to use the results of having executed error verification inquiries in the drafting of the analysis report and collection of multi-dimensional statistics, it (what is meant by “it”?) is composed of the modules that store the outcome of the error verification into the database.

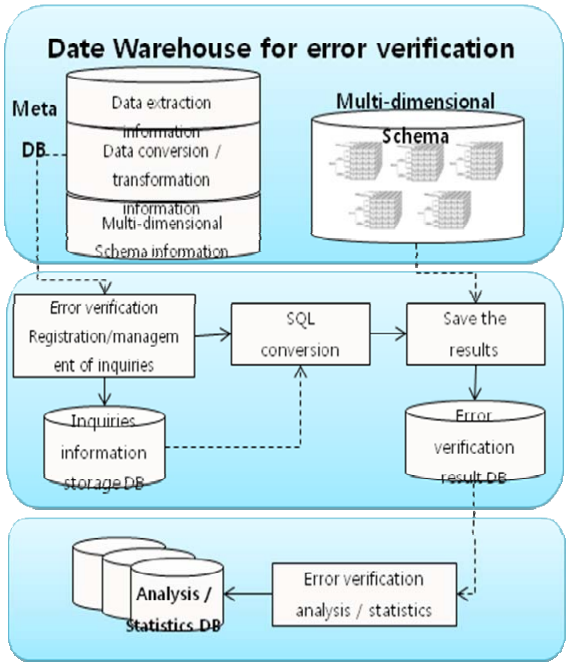


Figure. 4 Compositional diagram of the multi-dimensional error verification system module

- Management of multi-dimensional error verification inquiries: Manages the registration and setting of the multi-dimensional error verification inquiries.

- Relationship-type SQL conversion: Converts the multi-dimensional inquiries into the SQL inquiries used in the relationship-type database.
- Saving of the outcome of error verification: Stores and manages the outcome of the inquiries executed into a separate outcome database.

4. Conclusion

The error test is executed using S/W at the data test stage among the processes for establishment of database. The developed error test S/W cannot be used for error test at other institutions or business sites, and the S/W must be customized for application every time. Moreover, supplementation through development of relevant technology is necessary since there are extensive range of inadequacies in application including the problems of increase in the cost due to occurrence of large-scale correction and difficulties in securing of specialized development personnel. Although vast store of the database has been established through the project for establishment of the database executed by the national institutions, problems such as deterioration of the database quality and lack of compatibility of error test S/W arising from quantitative expansion of the database. To solve this problem, this study proposes a method for establishment of Data Warehouse for error verification, and design for the error verification system composed of modules for management of multi-dimensional error verification inquiries, conversion of relationship-type SQL, storing of outcome of error verification, and error verification analysis and statistics. For future research project, it is suggested that evaluation be conducted on the possibility of utilization of the designed error verification system developed as a prototype as a tool to enhance the quality of the database.

5. References

[1] Kim, Tae Hoon, “Study on Measurement and Evaluation of the Data Quality of the Data Warehouse” *The e-Business Studies Volume 7*, Number 2, June, 2006, pp. 417-444

[2] Data Quality Certification (DQC), http://www.dqc.or.kr/guideline/4-0-2.html

[3] Song, Hong Yul; Jeong, Gye Dong; Choi, Young Geun, “Designing of real-time data refining system for Data Warehouse by using XMDR”, *Journal of the Korea Ocean Research & Development Institute*, Book No. 14, Issue No. 8, 2010: pp. 1861-1867

[4] Jeong, Yong Wan; Jeong, Seung Guk, “Analysis of trend in the Data Warehouse(DW) appliance technology”, *Journal of the Korea Institute of Information Technology*, Book No. 10, Issue No. 2, 2012, pp. 107-111

Wireless Uroflowmetry System for Self Test at Home

In-Kwang Lee[1], A-Rong Heo[1], Ho Sun Shon[2], Keun Ho Ryu[2],
Kyoung-Ok Kim[3], Eun-Jong Cha[1], Kyung-Ah Kim[1]
[1]*Dept. of Biomedical Engineering, School of Medicine,*
Chungbuk National University, South Korea.
kwang4005@nate.com, coco2884876@naver.com, ejcha@chungbuk.ac.kr,
kimka@chungbuk.ac.kr
[2]*Database/Bioinformatics Laboratory, Department of computer science,*
Chungbuk National University, South Korea.
shon0621@dblab.chungbuk.ac.kr, khryu@dblab.chungbuk.ac.kr
[3]*Dept. of Nursing, Woosong College, Daejeon, Korea.*
kokim1989@hanmail.net

Abstract

Uroflowmetry is a non-invasive clinical test to diagnose benign prostate hyperplasia (BPH) frequent in aged men. Urine weight is measured in the standard method, while the present study suggests a new technique measuring pressure in toilet for self home care. Feasibility test was performed in simulated urination experiment followed by experiments with 8 normal adults. A very high correlation coefficient of 0.99998 close to an ideal value of 1 was obtained between volume and pressure in both wired and wireless transmissions. Volume measurement error was within 5% in average, demonstrating that accurate measurement was made during urination into toilet. Diagnostic parameters were evaluated from the volume signals in normal adults, all within the normal range. The present study verified feasibility of the wireless uroflowmetry on in-house toilet, which will be useful for self test and management of BPH.

1. Introduction

BPH is an important disease in aged societies, in particular, seriously causing problems in the quality of life [1-7]. Uroflowmetry may be the most popular screening test for BPH diagnosis owing to its non-invasiveness as well as simplicity[8], measuring the urine flow rate signal followed by evaluation of various diagnostic parameters. The standard technique measures changes in urine weight with a load cell, but resulting in significant impulsive noise generated on the bottom of the urine container. The authors previously suggested a new pressure measuring technique to minimize noise [9, 10]. The present study implemented a wireless uroflowmetry on toilet for convenient self test at home.

2. Materials and methods

2.1 Measurement principle

As shown in Figure. 1, pressure (P) accumulated while urination into a container with bottom area of A is proportional to the level (h), and accordingly proportional to the volume (V) of urine.

$$P=\rho gh=\rho gV/A \propto V \tag{1}$$

where ρ and g are density and gravitational constant, respectively. By definition, flow rate (F) is obtained by taking a time (t) derivative of V.

$$F(t)=\frac{dV(t)}{dt}=\frac{A}{\rho g}\cdot\frac{dP(t)}{dt} \tag{2}$$

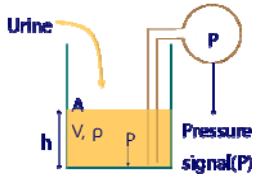


Figure 1. Uroflowmetry principle by pressure measurement

The urine container in Figure. 1 could be replaced with in-house toilet for self test as described below.

2.2 Experiment

Experimental set-up is depicted in Figure. 2. A catheter was dipped into toilet with appropriate curvature and connected to a pressure transducer (264, Setra, U. S. A.). Water was poured into toilet 10 times with each steps of 50 mL. P signal was acquired at end of each step, then transmitted to a personal computer (PC) in a wireless way with zigbee protocol, followed by evaluation of P-V relationship.

Simulated urination was performed by slowly pouring water of 250 mL with continuous acquisition of P in both wired (RS-232C) and wireless (zigbee) ways, simultaneously. V was compared for both transmission protocols. Experiment with 8 normal men with ages ranging 20-50 years was also performed after drinking 500 mL of water and waiting for enough time before voluntarily requiring urination.

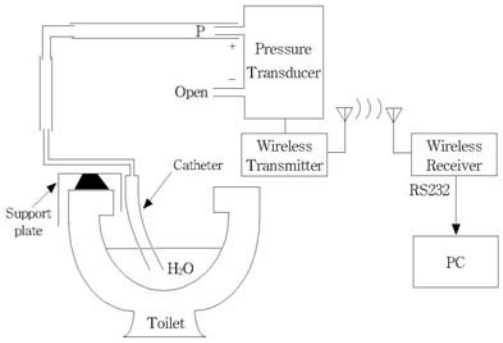


Figure 2. Experimental setup

3. Results

3.1 P-V relationship

P and V were mathematically fitted with a linear line as shown in Figure. 3. The last (10th) data point was significantly deviated from the line, thus eliminated in linear fitting. At this point, the curvature of toilet might have become significant, but V was 500 mL well higher than the usual volume during urination, justifying data elimination. The correlation coefficient was close to an ideal value of unity (0.99998), demonstrating that P measurement provided very much accurate V.

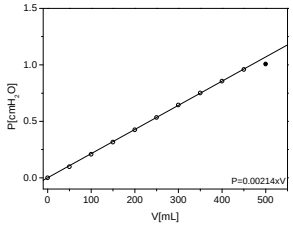


Figure 3. Relationship between V and P

3.2 Simulated urination

During simulated urination, the measured P signal was inserted into the P-V line shown in Figure 3. to obtain V signal. The results are presented in Figure 4. At end of simulated urination, the measured volume was 237 mL with approximately 5% relative error.

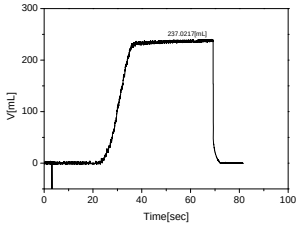


Figure 4. Volume signal example during simulated urination

3.3 Human experiment

During normal urination, V signal was obtained from P signal then differentiated to get F signal. A typical subject showed a very similar bell-shaped F signals in both wired and wireless transmissions as demonstrated in Figures. 5 and 6, respectively. Therefore, feasibility of the wireless uroflowmetry on toilet was verified appropriately.

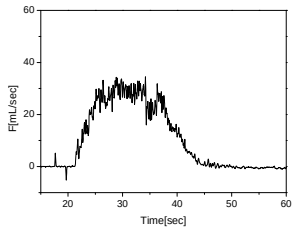


Figure 5. Uro flow rate signal (wired)

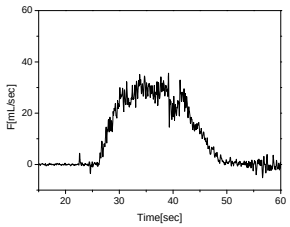


Figure 6. Uro flow rate signal (wireless)

4. Conclusions

The standard uroflowmetry measures urine weight with a load cell to obtain flow rate, which accompanies inconveniences of isolated urine container with measurement devices. Another problem may be significantly large impulsive noise generated on the container bottom. To prevent these problems, the present study implemented a new technique measuring pressure instead of weight and transmitting the signal in a wireless way, which makes possible convenient self test at home. Wireless transmission provides another advantage protecting the patient's privacy since the test can be self performed in bathroom with the door closed.

The mathematical relationship between volume and pressure obtained in the simulated urination experiment was a simple linear line with almost ideal correlation coefficient. The measurement error was within only 5%. Normal urination experiments demonstrated that the wireless transmission did not distort the flow waveform. Therefore, the feasibility of wireless uroflowmetry on toilet for self test at home was appropriately verified.

5. References

[1] B. Pradhan and K. Chandr, "Morphogenesis of nodular hyperplasia –prostate", *J Urol*, Vol. 13, 1975, pp. 210-219.

[2] P. Abrams, "In support of pressure-flow studies dor evaluating men with lower urinary symptoms", *Urology*, Vol. 44, 1994, pp. 153-155.

[3] E. Shapiro and H. Lepor, "Pathophysiology of clinical benign prostatic hyperplasia", *Urol Clin North Am*, Vol. 22, 1995, pp. 285-90.

[4] S. Berry, D. Coffey, and P. Walsh, "The development of human benign prostate hyperplasia with age", *J Urol* , Vol. 132, 1984, pp. 474-479.

[5] Y. G. Na, " Textbook of voiding dysfunction and female urology" , *Ilsogak*, Seoul, 2003, pp. 321-327.

[6] M. K. Cheong, " Textbook of benign prostatic hyperplasia", *Ilsogak*, Seoul, 2004, pp. 95-104.

[7] H. Guess, H. Arrighi, E. Metter, and J Fozard, "Cumulative prevalence of prostatism matches the autopsy prevalence of benign prostatic hyperplasia", *Prostate*, Vol. 17, 1990, pp. 241-246,.

[8] M. W. Kim, " Textbook of benign prostatic hyperplasia", *Ilsogak*, Seoul, 2004, pp. 127-126.

[9] K. A. Kim, S. S. Choi, I. K. Lee et al., "Validation of urine volume evaluation by hydraulic pressure measurement", *J Biomed Eng Res*, Vol. 28, No. 4, 2007, pp. 577-584.

[10] S. S. Choi, D. W. Cho, K. A. Kim et al., "Urinary flowmetry technique based on hydraulic pressure measurement", The 34th Conf, Kosombe, 2006, P2-4.

Context Ontology based Mobile Information Retrieval

Mi Sug Gu[1], Ho Sun Shon[2], Keun Ho Ryu[1]
[1]*Database/BioInformatics Laboratory, Chungbuk National University*
[2]*Graduate School of Health Science Business Convergence, Chungbuk National University*
[1]*{gumisug, khryu}@dblab.chungbuk.ac.kr*
[2]*shon0621@gmail.com*

Abstract

Background: In modern society, thanks to the development of Ubiquitous and IoT(Internet of Things) techniques, a lot of users can be provided personalized information. Among a variety of Ubiquitous techniques, using context ontology and user's context information it is possible to provide personalized information which users request. In this paper, we propose the mobile information retrieval system which finds out the information that users want, constructing the context ontology using the context information.

Methods: To construct the mobile information retrieval system for providing the personalized information which users want, constructing the context ontology using the context data, we used tour data as context information in this paper. We used internet sites which provide tour information to construct tour ontology, extracting tour information. We extracted tour site information such as the address of tour site, famous restaurant related to tour site, accommodation information, traffic information, and so on. Using the extracted information we developed tour context ontology. To develop tour ontology, we represented it using the ontology language OWL (Web Ontology Language), RDF(Resource Description Framework), RDFS(Resource Description Framework Schema) and so on. Also we used protégé to construct tour context ontology.

Results: It is possible to provide the exact and rapid retrieval results to user's request using the tour context ontology. In modern society, because we have been living in the flood of information, we need the method that can provide the information consistent to the user request information exactly and rapidly. Using the method proposed in this paper, it is possible to provide the information that the user wants.

Conclusion: In this paper we represented the mobile information retrieval method that can provide the information which user wants exactly and rapidly, using the tour context ontology. The method can improve the information retrieval efficiency compared with the existing retrieval system.

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).

Protected health information for research on computer forensics

Yoonhwan Shin, Keun Ho Ryu
Database/Bioinformatics Laboratory, Department of computer science,
Chungbuk National University, South Korea
cskisa@naver.com, khryu@dblab.chungbuk.ac.kr

Abstract

Background: With the development of ICT skills in the field of telemedicine medical services are being launched. Beginning stages of yet, but if we get started telemedicine services a patient's medical information should be legally protected. Telemedicine in accordance with the patient's medical information is kept digital materials, as well as easy, because the copy is also difficult to distinguish the original and the copy. Therefore, any computer that acts as a mediator to investigate whether the facts prove that the protected health information is necessary techniques.

Methods: In this paper, the integrity of the patient's medical information regarding the disk or file replication pattern is analyzed. And legal analysis of the evidence available to the pattern analysis technique is proposed.

Results: The proposed digital analysis techniques applied to a patient's medical information is a pattern that can be analyzed hacking. Through analysis of the pattern of digital data that can receive the legal protection can be based on collected data.

Conclusion: In this paper, the remote medical condition of the patient digital medical information disclosure can be proposed for the analysis of patterns. And it is based on a pattern of legal protection for computer forensic analysis techniques are presented.

Evaluating the Impact of Design Patterns on Code Design using Object-Oriented Metrics

Batnyam Battulga, Purev Jamai, Naranchimeg Bold, Tamir Chuluunbaatar
Information and Computer Science Department
School of Engineering and Applied Sciences, National University of Mongolia
Ulaanbaatar, Mongolia
{bbatnyam, purev, naranchimeg, tamir_chuba}@num.edu.mn

Abstract

The backbone of any software system is its code design. In software engineering, the most popular design technique is the Object-oriented design, and when using design patterns it provides good solutions for common OO problems and improves code design quality. In order to assess software quality, software metrics appeared a powerful and effective technology.

This paper presents the results of an experimental study of OO Code design which is investigating the relation between design patterns and OO Code design metrics.

***Keywords*:** Software Quality; OO Code Design Metrics; Design Patterns; Code Design, Software tools

1. Introduction

In software engineering, one of the most influencing factors of software quality is the structure of software design. To assess software quality more quantitatively and objectively, software metrics appear to be a powerful and effective technology.

There are a lot of OO metrics have been proposed in the literature [1]. Chidamber and Kemerer [2] proposed one of the most prominent metrics, which are used design measures for OO systems focusing on class and class hierarchy. Other relevant and important metrics are MOOD (Metrics for Object-Oriented Design) metrics which are proposed by Abreu et al. [3]. These metrics aim to measure the OO design in terms of encapsulation, inheritance, polymorphism and coupling. Lorenz and Kidd [4] proposed eleven metrics focuses on size, inheritance, internals and externals. Briand et al. [5] proposed metrics which are the measurement of the coupling between classes.

On the other hand, software design patterns are one of the most important topics that are language independent strategies for solving common OO design problems. Design pattern guarantees reusable design by avoiding creating objects directly, avoiding dependencies on specific operations, avoiding algorithmic dependencies and avoiding tight coupling [6].

The most famous collection of design patterns was presented in the book Design Patterns: Elements of Reusable Object-Oriented Software [7] where the authors have been known as Gang of Four (GoF). We used GoF design patterns in our experimental work.

Tuna Turk [8] investigated the relationship between the usage of software design patterns and software maintainability, as measured by MOOD and Chidamber and Kemerer metrics. Tuna Turk used Accepter Connector Pattern and Reactor Pattern, Smart Pointer Pattern, Forwarder Receiver Pattern and Command pattern on his/her master thesis.

Brain Huston [9] presented the analysis of the effect when applying various patterns on coupling, inheritance and method counts metric scores. The analysis undertaken has considered the three patterns Mediator, Bridge and Visitor.

The aim of this paper is to study the relation between design patterns and OO code design metrics, and review the improvements of software design using the value of design metrics after applying design patterns.

2. Software Design metrics and Design Patterns

2.1. Suit of Metrics for OO Design

In this study, we used the following suite of metrics.

The Chidamber and Kemerer metrics

The most prominent metrics of OO design are listed by Chidamber and Kemerer[2]. The following metrics used in this paper:

- Weighted Methods Per Class (WMC)
- Depth of Inheritance Tree (DIT)
- Number of Children (NOC)

CK metrics plays key role to know design aspects of the software and enhance the quality of software.

The MOOD metrics

The metrics of OO Design (MOOD) [3] set includes following metrics:

- Method Hiding Factor (MHF)
- Attribute Hiding Factor (AHF)
- Method Inheritance Factor (MIF)
- Attribute Inheritance Factor (AIF)
- Polymorphism Factor (PF)

MOOD metrics were defined to measure the use of OO design mechanisms such as inheritance (MIF and AIF) metrics, information hiding (MHF and AHF metrics), and polymorphism (PF metric).

The Lorenz and Kidd metrics

Lorenz and Kidd [4] proposed eleven metrics, the following OO metrics used in this experimental work:

- Number of Public Instance Methods (PIM)
- Number of Instance Methods (NIM)
- Number of Instance Variables (NIV)
- Number of Class Methods (NCM)
- Number of Class Variables (NCV)
- Number of Methods Overriden by a subclass (NMO)
- Number of Methods Inherited by a subclass (NMI)
- Number of Methods defined in a subclass (NMA)
- Class size metrics (Number of Class-NumClass, Number of Operations - NumOps, Number of Attributes - NumAttr)

Lorenz and Kidd metrics were defined to measure the static characteristics of software design, such as the usage of inheritance, the amount of responsibilities in a class.

Briand et al.'s metrics

Briand et al. [5] proposed metrics which are the measurement of the coupling between classes. We used following metrics:

- Number of times the class is externally used as attribute type (EC_Attr)
- Number of attributes in the class having another class of interface as their type (IC_Attr)
- Number of times the class is externally used as parameter type (EC_Par)
- Number of parameters in the class having another class of interface as their type (IC_Par)

2.2 Software Design Patterns

According to the [7], design patterns classified into three categories:

- Creational patterns, which deal with the process of object creation.
- Structural patterns, which deal primarily with the static composition and structure of classes and objects.
- Behavioral patterns, which deal primarily with dynamic interaction among classes and objects.

In this experimental work, we used Builder design pattern from Creational pattern category, Adapter design pattern from Structural pattern category and Chain of Responsibility, State Design Pattern from Behavioral pattern category.

The Builder pattern - is intended to separate the construction of a complex object from its representation so that the same construction process can create different representations.

Adapter design pattern - is intended to convert the interface of a class into another interface clients expect. Adapter lets classes work together that couldn't otherwise because of incompatible interfaces.

Chain of Responsibility pattern - is intended to avoid coupling the sender of a request to its receiver by giving more than one object a chance to handle the request. Chain the receiving objects and pass the request along the chain until an object handles it.

State design pattern - is intended to allow an object to alter its behavior when it's internal state changes. The object will appear to change its class.

3. Experimental Work

Evaluating the impact of design patterns on OO code design mechanism includes following processes:

1. ***Implement example projects*** – In order to calculate the metric values, we collected two source

codes (in Java) which are before applying design pattern and after applying design pattern for each design patterns.

2. Calculate metrics – To achieve the goal of metrics calculation of source code, we used Eclipse plug-in [11] for Metrics which calculates the metrics value from source files. To know that the calculated metrics values are correct, we verified it by the OO design quality measurement tool for the UML named SD metrics [10].

Figure 1 provides an overview of the tools and processes developed.

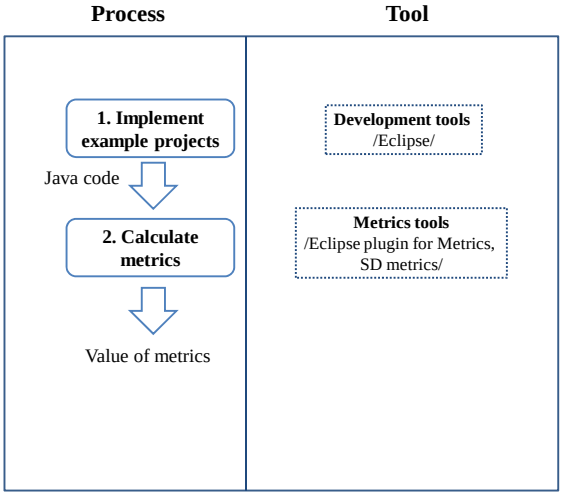


Figure 1. Tools and Processes used

4. Analysis of Experimental Results

The metrics chosen for analysis can be divided into 6 categories: Size, Inheritance, Coupling, CK metrics, Mood metrics, Lorenz and Kidd metrics as shown in the following tables. The shaded part represents the metrics value did not change after applying design patterns.

Table 1. Evaluating Results for Size Metrics

Design Patterns	Using DP	Size Metrics		
		NumClass	*NumAttr*	*NumOps*
Builder Design Pattern	Before	4	3	6
	After	5	10	14
Adapter Design Pattern	Before	3	0	2
	After	5	2	4
Chain of Responsibility	Before	2	1	1
	After	2	2	3
State Design Pattern	Before	2	1	2
	After	6	1	7

Table 2. Evaluating Results for Inheritance Metrics

Design Patterns	Using DP	Inheritance	
		NOC	*DIT*
Builder Design Pattern	Before	0	1
	After	0	1
Adapter Design Pattern	Before	0	1
	After	0	1
Chain of Responsibility	Before	0	1
	After	0	1
State Design Pattern	Before	0	1
	After	0	1

Table 3. Evaluating Results for Coupling (Briand) Metrics

Design Patterns	Using DP	Coupling Metrics			
		EC_Attr	*IC_Attr*	*EC_Par*	*IC_Par*
Builder Design Pattern	Before	N	N	N	N
	After	N	N	N	N
Adapter Design Pattern	Before	N	N	N	N
	After	N	N	N	N
Chain of Responsibility	Before	N	N	N	N
	After	N	N	N	N
State Design Pattern	Before	N	N	N	N
	After	N	N	N	N

Table 4. Evaluating Results for CK Metrics

Design Patterns	Using DP	CK Metrics		
		WMC	*DIT*	*NOC*
Builder Design Pattern	Before	23	1	0
	After	22	1	0
Adapter Design Pattern	Before	6	1	0
	After	6	1	0
Chain of Responsibility	Before	5	1	0
	After	8	1	0
State Design Pattern	Before	9	1	0
	After	11	1	0

Table 5. Evaluating Results for Mood Metrics

Design Patterns	Using DP	Mood Metrics				
		MHF	*AHF*	*MIF*	*AIF*	*PF*
Builder Design Pattern	Before	N	N	N	N	N
	After	N	N	N	N	N
Adapter Design Pattern	Before	N	N	N	N	N
	After	N	N	N	N	N
Chain of Responsibility	Before	N	N	N	N	N
	After	N	N	N	N	N
State Design Pattern	Before	N	N	N	N	N
	After	N	N	N	N	N

Table 6. Evaluating Results for Lorenz and Kidd Metrics

Design Patterns	Using DP	Lorenz and Kidd Metrics							
		PIM	NIM	NIV	NCM	NCV	NMO	NMI	NMA
Builder Design Pattern	Before	N	6	3	2	0	0	N	N
	After	N	14	10	1	0	0	N	N
Adapter Design Pattern	Before	N	2	0	1	0	0	N	N
	After	N	4	2	1	0	0	N	N
Chain of	Before	N	1	1	1	2	0	N	N

Responsibility	After	N	3	2	1	2	0	N	N
State Design Pattern	Before	N	2	1	2	0	0	N	N
	After	N	7	1	2	0	0	N	N

Followings are the observations made from applying the design patterns on samples:

- NumAttr is increased for all Design pattern except of State pattern, and both NumOps, NumClass are increased for all Design pattern except of Chain of Responsibility. This change indicates the functionality is improved.
- The increase of WMC in Design patterns Chain of Responsibility and State indicate the classes are little complex, therefore complexity of classes increased a little.
- The increase of NIM, NIV except of State pattern and zero value of NCV instead of Chain of Responsibility, shows the information hiding mechanism implemented well.
- NMO value is zero for all patterns. This shows the increase of cohesion and decrease of coupling.
- The values of PIM, NMA, NMI, MOOD Metrics and Briand Metrics are not direct calculated with eclipse plug-in [11] we are used.

5. Conclusions and Future Works

In this paper, we have taken and experimented few small sized examples for design pattern to keep effort to a minimum. However, we reached good usable results. In order to achieve a good result, we did not use all metrics because the value of WMC and NumOps, Coupling metrics and Briand metrics were the same.

For all Design Patterns, good encapsulation is achieved by better reusability and functionality, but complexity increases a little. Especially in State design pattern, we observed reusability increased meaningly in comparison to the design pattern unused version.

In this paper we experimented only for class source codes, in order to get more precise and good results we plan to experiment in another levels such as package, project, etc in the future.

6. References

[1] M. Xenos and D. Stavrinoudis and K. Zikouli and D. Christodoulakis, "Object-oriented metrics – a survey", proceedings of the ESMA 2000, *Federation of European Software Measurement Associations*, 2000, Madrid, Spain.

[2] Chidamber, S. and Chris Kemerer, F., A Metrics Suite for Object Oriented Design, *IEEE Transaction on Software Engineering*, 1994, Vol. SE-20.no.6, 476-493.

[3] Abreu, B. Fernando, Rita, Miguel, G.: "The Design of Eiffel Program: Quantitative Evaluation Using the MOOD metrics", *Proceeding of TOOLS"96 USA, Santa Barbara, California*, July 1996.

[4] M.Lorenz, and J.Kidd, Object-Oriented Software Metrics, *Prentice-Hall*, 1994.

[5] Briand L., Devanbu W. and Melo W.: "An investigation into coupling measures for C++", *19th International Conference on Software Engineering (ICSE 97)*, Boston, USA, 1997, pp. 412-421.

[6] Md. Abul Khaer, M.M.A. Hashem, Md. Raihan Masud, "On Use of Design Patterns in Empirical Assessment of Software Design Quality", *Proceedings of the International Conference on Computer and Communication Engineering*, 2008, p.133-137.

[7] E. Gamma, R. Helm, R.Johnson, and J. Vlissides. Design Patterns: Elements of Reusable Object Oriented Software. *Addison-Wesley Publishing Company,* Reading, MA, 1995.

[8] Tuna Turk, "The Effect of Software Design Patterns on Object-Oriented Software Quality and Maintainability", master thesis, 2009. www.eee.metu.edu.tr/~bilgen/TTurk.pdf

[9] Brain Huston, "The Effects of Design Pattern Application on Metric Scores", *The journal of Systems and Software*, 2001, p.261-269.

[10]Software Design Metrics Tool for the UML, http://www.sdmetrics.com/.

[11] Eclipse plug-in, http://metrics.sourceforge.net/

3D Reconstruction from uncalibrated images

Tsetsegjargal Erdenebaatar, Suvdaa Batsuuri
School of Engineering and Applied Sciences, National University of Mongolia
tsetsegee89@yahoo.com, suvdaa@num.edu.mn

Abstract

Reducing massive work of modeling environments of architect is interesting problem in computer vision. In this paper, we developed a method of this topic covering the 3d reconstruction process. People use methods reconstructing buildings that are manually modeling, 3D scanning and reconstructing from photographs. 3D reconstruction from uncalibrated image is one of the most fundamental and extensively researched topics in computer vision. The reconstruction of a building from uncalibrated image of scene might ease process of modeling 3D building.

Polynomial Approximation of Impedance of Microstrip Patch Antenna

Batpurev Mongol, Gerelmaa Byambatsogt, Ganbat Baasantseren
School of Information Technology National University of Mongolia
Ulaanbaatar, Mongolia
batpurev@num.edu.mn, gerlee_folo@yahoo.com, ganbatb@num.edu.mn

Abstract

In order to improve the radiation of the micristrip patch antenna, the impedance of the antenna must match with the impedance of the feeding cable. Hence, the impedance of the feeding cable is constant, we have to find the points on the microstrip patch antenna with a fixed impedance of feeding cable. To find the explicit form of the dependence of impedance from the position of the antenna is a difficult task. Instead, by measuring on the certain position the values of the function of impedance, we can approximate it on entire antenna. For this purpose, we used Lagrange and Taylor polynomial in this paper. From the results, we made a conclusion, that the approximation by Lagrange polynomials was close enough from direct measurement, and the difference did not exceed 5 percent.

Keywords*: Approximation; Lagrange polynomials; patch antenna, impedance.*

1. Introduction

Now, a microstrip antenna is widely used because the microstrip antennas are low profile, conformable to planar and nonplanar surface, simple, inexpensive to manufacture using modern printed-circuit technology and mechanically robust when mounted on a rigid surface. Major disadvantages are their low efficiency, low power, poor polarization purity. For identical antennas the directivity, S parameter (the reflection coefficient) is fundamentally different, depending from the feeding point. S parameter impact directly on the radiation of the antenna, i.e. the antennas with a poor S parameter, there will be no radiation. In order to increase efficiency, impedance of the antenna has to be matched with the feeder impedance. Hence the feeder impedance is constant and fixed, we have to find the points on the microstrip patch antenna with given feeder impedance. But the find the explicit form of the impedance of microstrip patch antenna is a difficult task. Instead, we can obtain the values of impedance by direct measurement on the fixed points. In other words, by dividing the antenna into equal intervals and measuring the impedance, we get the empirical information about impedance, varies over the surface of the antenna. But if we go over the interval of measurements in step, we can miss the optimal coordinates of the feeding point between intervals. Therefore, it is essential to know the explicit form of the impedance function, or approximate it by union of elementary functions. Our task is by table of the values finding the unknown form of function of impedance. In order to do this, in this paper, we used mathematical approximation by several series of polynomials, and choose the best appropriate series with less error. In this paper, to measure the value of impedance at given coordinates with given intervals, we used the CST Studio. In approximation, we used Taylor and Lagrange polynomials. The problem reduces to the fact that from the given empirical values find the coefficient of the both polynomials, and check the difference between the value of the approximated function with value of direct measurement. From the results, the Lagrange polynomials were better approximating to the seeking unknown function, than the Taylor series for the rim of the antenna. The difference between sixth or higher order was less than 1% of the value of measurement. The difference of the direct measurement of the antenna at given point and value of the Lagrange polynomials at that point, for the patch antenna wasn't exceeding the 5% of its value.

2. S parameter

S parameter indicates the interrelation between input and output ports of the electrical system. Figure 1 shows the general block diagram of the electrical system with two input ports.

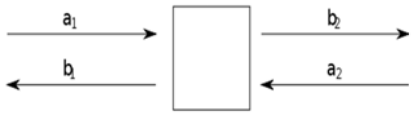


Figure 1. Block diagram of electrical system with two input ports

Expression (1) shows the interrelation between receiving and reflected power by S matrix:

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad (1)$$

Here a1 indicates the signal power in coming to the first input, b1-outcoming signal power from the first input, a2-signal power in coming to the second input, b2-outcoming signal power from second input, S12- power, transmitted from input 2 to input 1, S21-power transmitted from input 1 to input 2. S22-reflected power from the second input while S11-reflected power from first input.

In this case, our electrical system has only one input and S matrix is a scalar. This scalar is denoted S11. S11 parameter indicates a portion of power, that is reflected from the antenna and it is called the reflection coefficient. In some cases it is called reflected loss and denoted by Γ. Therefore, if S11=0, then all the power is reflected from the antenna, and nothing has radiated from the antenna.

3. Approximation Methods

In solving many problems often can be obtained only empirical information about the processes and phenomena. Typically, this information is a table of values of the unknown function. Find the explicit form of this function – practically significant. We can choose a different, more simple function F (x), in some sense, is close to the unknown function of the process. In addition, other construction of functions is used, in cases where the values of the calculation are very difficult and requires less accuracy. With the selected function, you can calculate the approximate values that are not contained in the table.

The general problem is that we have the table of the function:

Table 1. Values of the approximating function

x	x_0	x_1	…	x_k	…	x_n
$y = f(x)$	y_0	y_1	…	y_k	…	y_n

Required to construct a function F(x), that takes at the points x_k, $k = \overline{0,n}$ the same values as the unknown function, i.e. the equality holds (basic property of interpolation), $F(x_k) = f(x_k) = y_k$, $k = \overline{0,n}$.

The function F(x) is called interpolating function, and points $x_1, x_1 \ldots x_n$ are interpolation nodes.

Geometrically interpolation means, that you need to find a curve y=F(x), passing through a given a system of points $M(x_k, y_k), k = \overline{0,n}$.

We seek a polynomials $P_n(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_1 x + a_0$, such that $P_n(x_i) = f(x_i) = y_i$, $i = \overline{0,n}$.

If we expand the latest, the we get the equation:

$$\begin{cases} P_n(x_0) = a_n x_0^n + a_{n-1} x_0^{n-1} + \ldots + a_1 x_0 + a_0 = y_0, \\ \cdots \\ P_n(x_n) = a_n x_n^n + a_{n-1} x_n^{n-1} + \ldots + a_1 x_n + a_0 = y_n. \end{cases} \quad (2)$$

by introducing the notation

$$X = \begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ \cdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \quad (3)$$

$$a = (a_0, \ldots, a_n)^T \quad (4)$$

$$y = (y_0, \ldots, y_n)^T \quad (5)$$

then the system can be written in matrix form:

$$Xa = y \quad (6)$$

To find the coefficients of the polynomials, we have to solve the matrix equation above.

Another way of interpolation is Lagrange polynomials. Let the function $f(x)$ on the interval $[a,b] = [x_0, x_n]$ is specified table of its values $[x_i, y_i = f(x_i)], i = \overline{0,n}$. Required to construct a polynomial $L_n(x)$ of degree n, satisfying the

conditions $L_n(x) = y_i, i = \overline{0,n}$. Such a polynomial is sought in the form:

$$L_n(x) = \sum_{i=0}^{n} l_i(x) \tag{7}$$

where $l_i(x)$ is a polynomial of degree n, according to the basic property of the interpolation must satisfy the conditions:

$$l_i(x_k) = \begin{cases} y_i & , \ i = k, \\ 0 & , \ i \neq k. \end{cases} \tag{8}$$

We will seek it in the form:

$$l_i(x) = c_i(x - x_0) \times \ldots \times (x - x_{i-1}) \times$$

$$\times (x - x_{i+1}) \times \ldots \times (x - x_n) \tag{9}$$

where the coefficients c_i to be determined. After a simple calculation, we find the coefficients c_i, substituting it back, we get a Lagrange polynomial:

$$L_n(x) = \sum_{i=0}^{n} y_i \frac{\prod_{n+1}(x)}{(x - x_i)\prod_{n+1}'(x_i)} \tag{10}$$

In the case of equidistant nodes, with a constant pitch

$h = x_{k+1} - x_k = Const$ the Lagrange polynomial has the form:

$$L_n(x_0 + th) = \sum_{i=0}^{n} y_i \frac{(-1)^{n-i} \mathrm{T}_{n+1}}{i!(n-i)!(t-i)} \tag{11}$$

4. Simulation

To simulate, the several models of the microstrip patch antenna, are used. The impedance, the radiation power and the resonant frequencies depends on feeding point. Therefore, finding the optimal feeding point is typical subject to be studied, to improve antenna parameters.

We seek to find the optimal feeding point of the microstrip patch antenna to improve the efficiency of radiation. To make a simulation and calculate the input impedance, S11 parameter depending on feeding point, we used CST Microwave Studio.

We proved that the antenna has symmetry in respect to input impedance and S11 parameter [5]. So we share the antenna into 4 symmetrical quarters, and the calculations were made only in the first quarter. By dividing the length in step of 2mm, and width - 3.1mm, we made a 56 measurements. And we had 56 impedance value of feeding position.

Figure 2a demonstrates the relation between input impedance magnitude and feeding point in 3D. Figure 2b shows upside of Figure2a. The graph shows, that minimum of the impedance is in the center of antenna (5.88 Om) and increases at the boundary (94.4 Om). Note there was 195 coordinates because we mirror reflected into other quarters.

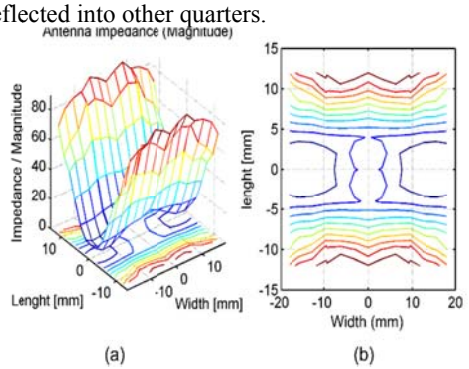


Figure 2. Input impedance magnitude vs feeding point

5. Approximation of Simulation

We have impedance values. Now we need to calculate impedance of arbitrary feeding point. We will use specify the function using a table of values. We use two methods and compare them: Taylor series and Lagrange polynomial.

The first, we found coefficients of 3rd order Taylor polynomials when x position of feeding point was fixed. And wrote 3rd order Taylor polynomials. Table 2 shows the 3rd order Taylor polynomial coefficients.

Table 2. 3rd order Taylor polynomial coefficients

Position of x (mm)	3rd order Taylor series coefficients			
	y^3	y^2	y^1	y^0
0	0.1812	-0.8287	0.3975	23.37
3.1	0.6321	-4.9375	9.4367	21.5
6.2	0.1394	- 0.0525	– 1.1025	16.4
9.3	-0.3	4.2358	– 10.1533	10.93
9.5	0.0735	0.51	– 1.1692	10.57
12.6	-0.025	– 1.1845	– 1.6086	5.99
15.7	-0.0471	1.3863	– 2.1392	5.83
18	-0.042	0.9088	– 1.6508	8.9

Then, we found coefficients of 6th order Taylor polynomials when the x position of feeding point was fixed. And write 6th order Taylor polynomials. Table 3 shows the 6th order Taylor polynomials coefficients.

Table 3. 6th order taylor polynomials coefficients

x (mm)	6th order Taylor series coefficients						
	y^6	y^5	y^4	y^3	y^2	y	y^0
0	-0.0016	0.0054	-0.7025	4.3522	-11.8794	10.6261	23.37
3.1	-0.0011	0.0391	-0.5439	3.5332	-9.3433	7.644	16.4
6.2	-0.0011	0.0391	-0.5439	3.5332	-9.3433	7.644	16.6
9.3	-0.0001	0.064	-0.1163	0.9379	2.1311	1.4687	10.93
9.5	-0.0006	0.0218	-0.3034	1.9762	-4.7242	3.768	10.57
12.6	0.0002	-0.0041	0.0279	-0.0722	1.094	-1.4018	5.99
15.7	0.001	-0.031	0.3616	-1.9908	6.1580	-6.336	5.83
18	0.0006	-0.0191	0.2005	-0.968	3.0452	-3.3882	8.9

In Lagrange, we use 5.1 expressions:

$$L_n(2t) = \sum_{i=0}^{6} Z_i \frac{(-1)^{6-i} \mathrm{T}_7}{i!(6-i)!(t-i)} \qquad (12)$$

Last, we checked our expression by CST Microwave Studio. Table 4 presents a comparison of the Taylor polynomials result and Lagrange polynomial's result. Table 5 shows their errors in percentage.

Table 4. Comparison of the Taylor polynomials and Lagrange polynomials's results

Feed position		Taylor polynomials		Lagrange	Simulation result /Ω/
x /mm/	*y /mm/*	*3rd order /Ω/*	*6th order /Ω/*		
0	1	23.48	25.81	24.37	22.90
	5	27.65	25.20	28.06	27.85
	9	91.91	58.71	62.74	63.10
3.1	9	167.29	78.99	68.03	68.90
6.2	1	15.38	17.74	17.74	15.97
	9	107.67	59.97	70.8	72.85
9.3	1	4.7	11.09	11.09	10.80
	9	43.95	96.98	72.04	72.99
9.5	1	9.98	11.32	11.32	10.54
	9	94.93	80.42	72.85	72.78
12.6	1	5.54	5.63	5.63	6.07
	9	69.23	76.59	60.43	60.37
15.7	1	5.03	3.99	3.99	5.7
	9	64.53	69.69	58.15	58.86
18	1	8.11	7.77	7.77	8.7
	9	37.03	25.90	51.69	53.11

Table 5. Comparison of errors in percentage

Feed position		Taylor polynomials		Lagrange
x /mm/	*y /mm/*	*3rd order*	*6th order*	
0	1	2.53%	12.70%	6.4%
	5	1.72%	9.52%	0.78%
	9	45.65%	6.95%	0.57%

Feed position		Taylor polynomials		Lagrange
x /mm/	*y /mm/*	*3rd order*	*6th order*	
3.1	9	143%	14.60%	0.86%
6.2	1	3.60%	11%	11%
	9	47.80%	17.65%	2.78%
9.3	1	56.50%	2.68%	2.68%
	9	39.79%	32.86%	1.31%
9.5	1	4.50%	8.32%	8.32%
	9	30.43%	10.49%	0.01%
12.6	1	8.74%	7.25%	7.25%
	9	14.70%	26.86%	0.09%
15.7	1	11.76%	30%	30%
	9	9.63%	18.39%	1.21%
18	1	6.90%	10.68%	10.68%
	9	30.28%	51.24%	2.68%

6. Conclusion

In order to improve the radiation of the antenna, the impedance of the antenna must match with the impedance of the cable. The antenna impedance of varies with the location of the feeding point. Therefore, finding the dependence of the impedance of feeding point, the problem is almost significant, but difficult to achieve. Instead of specifying impedance functions from feeding points, approximating it based on some measured points much faster. In this paper, we used approximation by two types of polynomials. First is a Taylor polynomials, and the second is Lagrange polynomials. In Table IV we see that Lagrange polynomials approximately match simulation results. In Table V we see, that Lagrange polynomial's maximum error is 11% and minimum error is 0.01%. As we see from the results Lagrange polynomials better than Taylor polynomials.

7. References

[1] G.A Deschamps, "Microstrip microwave antenna," Presented at the Third USAF Symposium on Antenna, 1953.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[2] Талагаев Ю.В., Тараканов А.Ф. Методы аппроксимации и обработки экспериментальных данных в среде MathCad. Учебное пособие / Ю.В.Талагаев, А.Ф.Тараканов. – Борисоглебск, 2008. – с.: ил.

[3] C.A. Balanis, "Antenna Theory Analysis and Design", Third edition, 2005.

[4] C.A. Balanis, "Advanced Engineering Electromagnetics", New York, John Wiley&Sons, 1989.

[5] Inder J.Bahl, "Microstip Antennas", First edition

[6] G. Byambatsogt, B.Mongol, G. Baasantseren "Feeding Point Analysis of Microstrip Patch Antenna", The 6th International Conference FITAT 2013 and 3rd International Symposium ISPM 2013

[7] Njeri P. Waweru, D.B.O. Konditi, and P.K. Langat, "Variation of Input Impedance with Feeding Position in Probe and inset-Fed Microstrip Patch Antenna", *Innovative Systems Design and Engineering*, ISSN 2222-1727 (Paper) ISSN 2222-2871 (Online), Vol 3, No 7, 2012.

[8] T. Samaras, A. Kouloglou, and J.N. Sahalos "A Note on the Impedance Variation with Feed Position of a Rectangular Microstrip-Patch Antenna", *IEEE Antennas and Propagation Magazine*, Vol. 46, No.2, April 2004.

[9] W.F. Richards, Y.T. Lo, and D.D. Harrison "An Improved Theory for Microstrip Antennas and Applications", *IEEE Transactions on Antennas and Propagation*, AP-29, 1, 1981, pp. 38-46.

[10] J.R. Mosig, R.C. Hall, and F.E. Gardiol "Numerical Analysis fo Microstrip Patch Antenna", in J.R. James and P.S. Hall (eds.), *Handbook of Microstrip Antennas*, London, Peter Peregrinus Ltd., 1989

[11] L.I. Basilio, M.A. Khayat, J.T. Williams and S.A. Long, "The Dependence of the Input Impedance on Feed Position of Probe and Microstrip Line-Fed Patch Antennas", *IEEE Transactions on Antennas and Propagation*, AP-49,1, 2001, pp. 45-57.

[12] V.R. Anitha and S.N. Reddy "Design of an 8X1 square microstrip patch antenna array", *International journal of electronic engineering research*, vol.1, no.1, pp. 71-77.

[13] J.M.Rathod "Comparative study of microstrip patch antenna for wireless communication application", *International journal of innovative and technology*, vol.1, pp 194-197.

[14] M.A. Matin and A.I. Sayeed, "A design for inset fed rectangular microstrip patch antenna", *WSEAS transaction on communications*, vol.9, January, pp. 63-72.

Hand gesture controlled drawing tool using "Asus xtion pro"

Amartuvshin Renchin-Ochir, Dorjnamjirmaa Badraa
School of Applied Science and Engineering,
National University of Mongolia,Ulaanbaatar, Mongolia
{*amartuvshin.r, dorjoo.b*}*@gmail.com*

Abstract

Depth data acquired by current low-cost real-time depth cameras provide a very informative description of the hand pose, that can be effectively exploited for gesture recognition purposes. In this paper, we introduce recognition hand location and touch event sensing algorithm using a "asus xtion pro" depth camera. The hand is firstly extracted from the acquired depth maps with the aid also of color information from the associated views. Then the hand is segmented into palm and finger regions. The proposed we developing a hand gesture controlled drawing tool based on depth data.

Keywords*—Depth data; Hand controlled drawing tool*

1. Introduction

Hand gesture recognition and hand tracking systems are an intriguing problem that has many applications in different fields, such as human computer interaction, robotics, computer gaming, automatic sign-language interpretation and so on [1]. A certain number of hand gesture recognition approaches, based on the analysis of images[2, 3], but it is difficult and many problems. Bi dimensional representation is not always sufficient to capture the complex movements and inter-occlusions characterizing hand gestures. The recent introducing low cost depth camera is giving us big chances of analyses based depth information.

Sample images contains x and y two dimension and intensity, but depth images contains two dimensions and depth information (show in fig 1). Depth information is very important for recognition to hand location and gesture. Also the depth information can not only improve the hand location, but also enable the estimation of 3D hand positions instead of 2D positions [4]. The purpose of features is to ease the fast and accurate detection of the hand region in a depth image. Many researchers use very simple features and let the boosting and cascade method learn how to detect a hand region using the features [5].

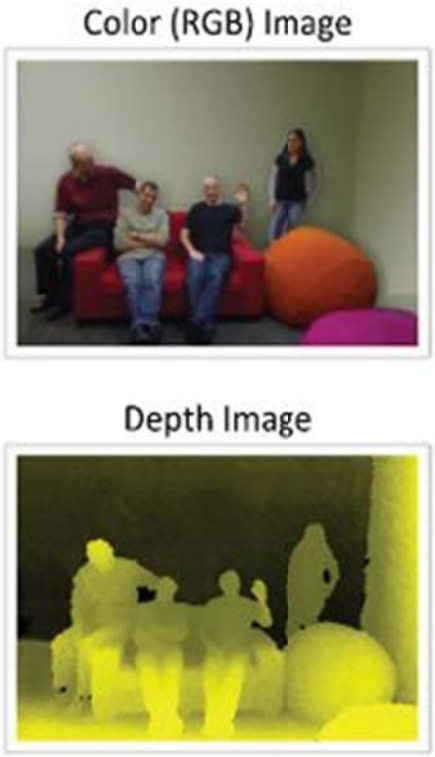


Figure 1. RGB images and depth images difference

The Gaussian mixture model is used for calculating the probability distribution of the 3D x-coordinates and then to detect the hand and the forearm regions. This method detects a static pose, but it is limited when used for dynamic gesture recognition because the distribution model needs to be revised when the depth data change. Suryanarayan et al. [6] proposed 2D figure data, a compressed 3D figure descriptor, and a 3D volume metric figure descriptor for hand pose recognition that uses depth data. The hand is detected by creating a histogram of depth values and the

detected hand is separated from others by Otsu's threshold method. This method is also limited when there is another object between the camera and the hand. Oikonomidis et al. [7] used a Kinect camera sensor [8] to detect a hand. It uses the hand model with all degrees of freedom. Then, it initializes the hand model with the hypothesized pose and keeps tracking a hand in real time by updating the hand model. This method optimizes the hand model parameters through minimizing the difference between the assumed hand model in the 3D space and the actual hand. However, it recognizes the hand pose by comparing adjacent distances; therefore, an error may occur because the hand pose becomes increasingly blurry with an increase in the distance [5].

The proposed "hand controlled drawing tool" consists of main two steps with hand detection and to draw. In this study, we use the depth data acquired from the "asus xtion pro" depth camera.

2. Hand detection

The proposed system detects hand click events and performs hand tracking on depth images. When a click event is detected, the system takes the 3D location of the event as the initial position and activate the hand tracking algorithm. In this section, the hand tracking algorithm and the method to detect click events are depicted. For each frame t, the input image takes the form of depth image $\mathcal{D}$ where each pixel indicates the distance from this point to the depth sensor. Since the intrinsic parameters of the depth sensor are known, the input depth $\mathcal{D}$ can be transformed into a 3D point cloud $\mathcal{P}$ containing 3D points with known $p = (x, y, z)$ coordinates.

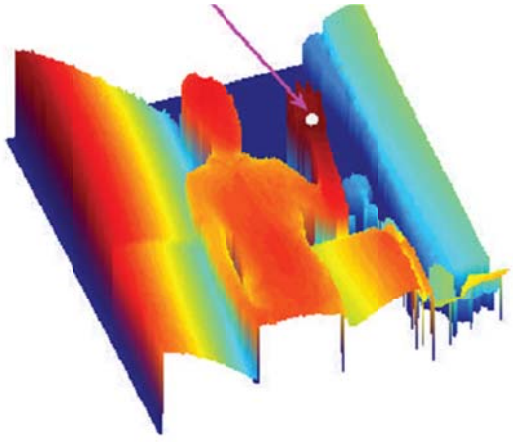


Figure 2. 3D plot of depth camera images

Given the hand position in the previous frame, the hand tracking algorithm finds the new hand position in the current frame. The proposed algorithm involves three stages. The first stage is hand position prediction based on the hand moving velocity. The second stage segments the entire hand region using a region growing technique on the 3D point cloud. Given the segmented hand region, the last stage estimates the new 3D position of the hand. The estimated position should be semantically consistent across frames because it serves as the output of the tracking algorithm, as well as the initial hand position for the next frame. (show in Fig 2). Given the previous hand position H_{t-1}, we predict the new hand position based on the hand moving velocity:

$$H_t^{pred} = H_{t-1} + v \qquad (1)$$

where v is the hand moving velocity estimated from hand positions tracked in previous frames.

The entire hand region can be found as a connected component in the point cloud P . To begin with, we first use the predicted hand position to find a seed point:

$$H_t^{seed} = \arg\min_{p \in P} d\left(p,\ H_t^{pred}\right) \qquad (2)$$

where d(·, ·) denotes the Euclidean distance between two points. The seed point H_t^{seed} is the nearest point in the point cloud P from the predicted hand position H_t^{seed}. As shown in Fig. 2 for white point, the predicted hand position and the seed point are indicated as red and blue circles respectively. The connectivity between two points p, q in the point cloud P can be defined based on Euclidean distance as follows:

$$connected(p,q) = \begin{cases} 1\ if\ d(p,q) < \delta \\ 0\ otherwise \end{cases} \qquad (3)$$

where δ is a pre-defined threshold specifying how far from each other two connected points can be. We use δ = 30 mm through our experiments.

Given the seed point H_t^{seed} and the connectivity defined in (3), the entire hand region can be segmented as a connected component using standard region growing techniques. Beginning at the seed point, the hand region grows incrementally and stops when two neighboring points are no longer connected, that is, their distance exceeds the specified threshold δ. The region growing also stops when the geodesic distance from the current point to the seed point is longer than τ , which is set as 250 mm so that the entire hand region can be found without including other body parts. We

denote the set of hand points found in this process as Ω, and the set of depth boundary points as ξ where the region growing stops because of depth discontinuity. Where Ω and ξ are shown in green and red respectively. Once the entire hand region is segmented, the remaining task is to localize the hand center. We first estimate a rough hand center according to the set of depth boundary points ξ:

$$H_t^{boundary} = \arg\max_{p\in P} \sum_{p\in\Omega} 1_{d(p,q)<r} \qquad (4)$$

where 1 statement = 1 when the statement is true and 0 otherwise. H boundary t is determined to be the point in Ω with maximum boundary points in its neighborhood. The size of the neighborhood is specified by r, which is 120 mm in our experiments. The intuition behind this operation is that there should be more boundary points around the palm than around the forearm.

Boundary points ξ are rendered in red. Then we use mean-shift to refine the hand center location and try to locate the very center of the palm. $H_t^{boundary}$ is used as the initial point in the mean-shift algorithm. The mean position of hand points inside its neighborhood is calculated as follows:

$$H_t^{mean} = \frac{\sum_{p\in\xi} p\cdot 1_{d\left(p,H_t^{boundary}\right)<r}}{\sum_{p\in\xi} 1_{d\left(p,H_t^{boundary}\right)<r}} \qquad (5)$$

Although this mean-shift procedure can be performed iteratively, we found that one iteration is good enough in our experiments. Note that after the mean-shift step, H mean t is not guaranteed to be one of the hand points Ω, since it is the mean position of all nearby hand points. So we finally determine the hand center as the point in Ω with the nearest viewing direction from the sensor:

$$H_t = \arg\max_{p\in\Omega} cos(p, H_t^{mean}) \qquad (6)$$

where $\cos(p, H_t^{mean})$ is the cosine of the angle between two vectors p and H_t^{mean} starting from the sensor, as illustrated in Fig. 3.

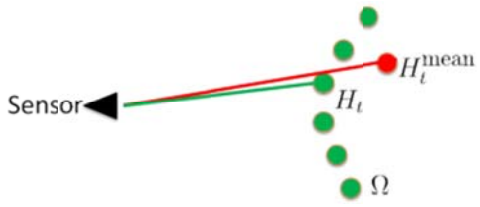


Figure 3. Hand center localization after one step mean shift

It can be seen that the hand center is located at similar semantic positions regardless of different hand shapes. The hand center position determined with the elaborated technique employed in this stage suffices for hand tracking algorithm output and avoids jittering across frames without smoothing filtering [4].

3. Experimental results

In this section, We developed of a sample drawing tools and we defined as H_t^{mean} composed by x, y and z, in previous section. So we can sensing for vertical and horizontol moving of the hand, and we can detect distance between the camera and hand.

.

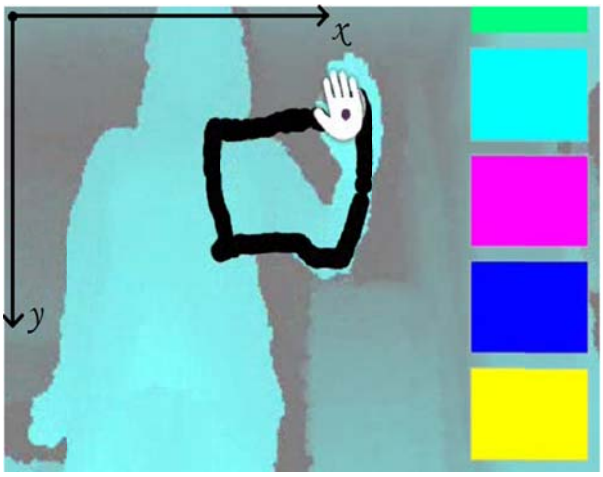


Figure 4. Sample form of the drawing tools

The sample drawing tool form contains two main sides they named for ‘color select’ and ‘main draw’. And tool is activation for few steps. They are:

- Check the main camera is “Asus xtion pro”.
- Wave gesture recognition then running of the drawing tools form (show in fig 4).
- Then determine of the hand position.
- Check touch event for distance between hand and camera. If touch event is sensing, drawing on the form. But else just moving cursor.
- Also, firstly selected black is main color and we can choose other colors on the ‘color select’ sides.
- Then, if we want stop the tool, we should shake my hand.

We show in fig 5,6 tested some results.

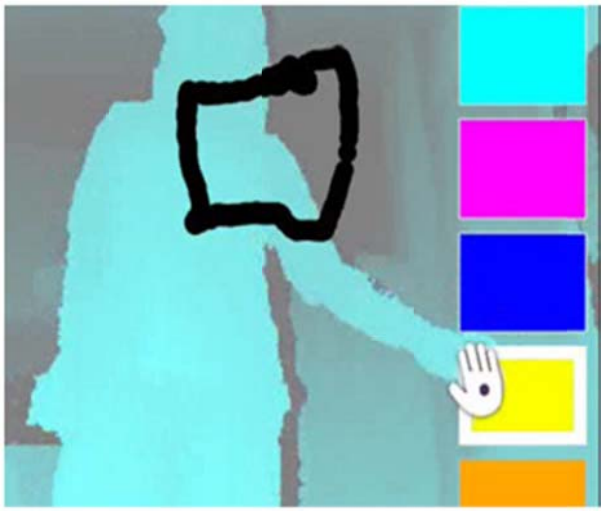


Figure 5. Color selecting and selected color is white borders

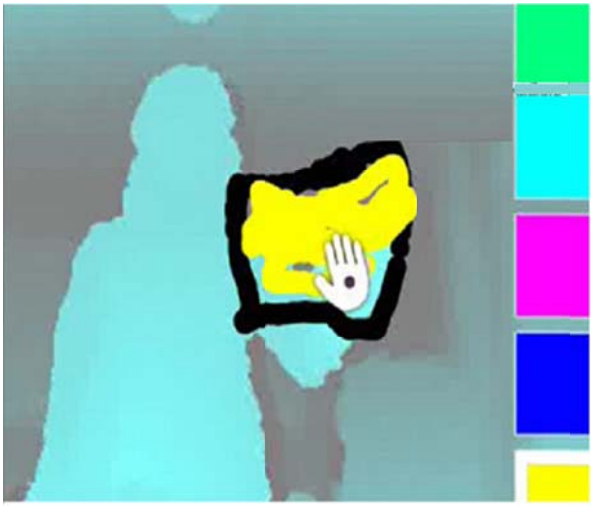


Figure 6. Drawing a yellow color.

4. Conclusion

In this paper, The proposed hand tracking algorithm enable of interface applications where human device interaction is based on hand gestures. The tracked hand positions can be used to control the cursor, and can be recognition hand gesture, such as touch, shake hands, or wave hands, can also be easily detected according to the tracked hand positions. Then we developed a sample drawing tool on C#. Also we using by "Asus xtion pro" depth camera.

5. References

[1] F. Dominio, M. Donadeo, G. Marin, P. Zanuttigh, G.M. Cortelazzo, "Hand Gesture Recognition with Depth Data", ACM Multimedia Artemis workshop 2013, Barcelona, Spain, October 2013.

[2] D. Kosmopoulos, A. Doulamis, and N. Doulamis. "Gesture-based video summarization. In Image", Processing, 2005. ICIP 2005. IEEE International Conference on, volume 3, 2005, pages III–1220–3.

[3] J. P. Wachs, M. K¨ olsch, H. Stern, and Y. Edan, "Vision-based hand-gesture applications", Commun. ACM, 54(2), Feb. 2011, pp:60–71.

[4] C.-P. Chen, C. Yu-Ting, L. Ping-Han, T. Yu-Pao, and L. Shawmin, "Real-time hand tracking on depth images," in Visual Communications and Image Processing (VCIP), 2011, pp. 1-4.

[5] Sung-Il Joo, Sun-Hee Weon, and Hyung-Il Choi, "Real-Time Depth-Based Hand Detection and Tracking", the Scientific World Journal Volume 2014, Article ID 284827, 2014, 17 pages.

[6] P. Suryanarayan, A. Subramanian, and D. Mandalapu, "Dynamic hand pose recognition using depth data," in Proceedings of the 20th International Conference on Pattern Recognition (ICPR '10), August 2010, pp. 3105–3108.

[7] I. Oikonomidis, N. Kyriazis, and A. A. Argyros, "Efficient model-based 3D tracking of hand articulations using Kinect," in Proceedings of the British Machine Vision Conference , Dundee, UK, 2011, pp.101.1–101.11.

[8] PrimeSensor,http://www.primesense.com/.

[9] E. Stergiopoulou and N. Papamarkos. Hand gesture recognition using a neural network shape fitting technique. Engineering Applications of Artificial Intelligence, 22(8), Dec. 2009, pp:1141–1158.

[10] Andrew Davison, "Techniques for Games.Kinect", Chapter 8. Hands Tracker, Draft #1 (14th Nov. 2011)

Analyzes of enrollment database of an University Information System

Bulganchimeg.B[1], Naranchimeg.B[1], Oyun-Erdene.N[1], Yanjindulam D[2], Sodbileg.Sh[1]
[1]*School of Engineering and Applied Sciences, National University of Mongolia, Ulaanbaatar, Mongolia*
{bulgaa, naranchimeg, oyunerdene, sdblg}@num.edu.mn
[2]*Department of Mathematics and Computer Science Eindhoven University of Technology, Eindhoven, Netherlands*
y.dajsuren@tue.nl

Abstract

SISi is web based software for effective management of administrative and university management functions that are necessary to successfully handle all of the challenges of running National University of Mongolia. This system consists of following modules and subsystems: Admission, E-Journal, Time table, Forum, OCW open course, Lecturer's web site, Curricula, Voting and Questionary, Course records, Contract agreement, Student records, Grade records, Information distribution, Online Testing, Course selection, Reporting, Finance, Human Resource. SISi system also interacts with Entrance Exam System (Education Evaluation System) Library System, Wi-Fi hotspot, E-Finance system, Student's card system, Voting System for Best Student award, National University of Mongolia's web site, Online payment systems.Once we have both enrollment database and Student record Subsystem, our goal is set to discover knowledge useful in understanding the university enrollment and the find ways to increase number of students and successful graduates. In the discovery process we have been used "Weka". Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is also free software available under the GNU General Public License. We have shown that good high school students are the best source of large numbers of credit hours and increases number of successful graduates, while some of students drop out, causing significant enrollment loses. We met many interesting findings and surprises, motivating us to expand our exploration. The discovered knowledge can affect decision making and policy formation at National University of Mongolia.

Improvement of the database performance of an University Information System

Munkhtuya.D, Naranchimeg.B, Oyun-Erdene.N, Sodbileg.Sh
School of Engineering and Applied Sciences, National University of Mongolia
{munkhtuya, naranchimeg, oyunerdene, sdblg}@num.edu.mn

Abstract

SISi is web based software for effective management of administrative and university management functions that are necessary to successfully handle all of the challenges of running National University of Mongolia. This system consists of following modules and subsystems: Admission, E-Journal, Time table, Forum, OCW open course, Lecturer's web site, Curricula, Voting and Questionary, Course records, Contract agreement, Student records, Grade records, Information distribution, Online Testing, Course selection, Reporting, Finance, Human Resource. SISi system also interacts with Entrance Exam System (Education Evaluation System) Library System, Wi-Fi hotspot, E-Finance system, Student's card system, Voting System for Best Student award, National University of Mongolia's web site, Online payment systems.We manually transferring data from above systems, therefore we need to build Real-time Data-Warehousing for SISi system. Currently our SISi system's database is located on MSSql database management system. Even if DB Server's performance is low as 16GB RAM, and with 80GB hard, size of database increasing academic year by year. We keeping historical data on backup devices, so cause of limited hardware design we can't mine huge amount of valuable data. And there is no any software based decision support system is working at NUM. In this research we demonstrated whole data flow of University management system, and represented Real Time Data Warehousing for SISi System. Our experimental results shows that, improvement of Database design, Real-Time Warehousing could increases not only SISi system's performance but also it's sub system's performances. Therefore we could mine historical huge amount of data of University management system and discovered findings can affect perfectly to decision making and policy formation at National University of Mongolia.

Quadrupeds Motion Data Collection Method

Javkhlan Rentsendorj, Erdenebat Budsuren, Baatarbileg Altangerel, Oyun-erdene Namsrai
School of Information Technology, National University of Mongolia
javkhlan@seas.num.edu.mn, {free3erkaa@, a_bbileg, oyun_erdene79}@yahoo.com

Abstract

Highly realistic, computer generated creatures creating technical challenges have been receiving attention recently. Lots of field in entertainment, medical and education industries have increased the demand for simulation of realistic animals in the computer graphics area. In order to achieve this we need to gather and process motion data that embodies from an animal. Most animals specially quadrupeds cannot be easily motion-captured and building accurate kinematic models for animals with adapted animation skeletons. Developing kinematic or physically-based animation methods needs to embedding some a prior knowledge about the way that quadrupeds locomotion adopting examples of real motion. In this paper, we present an overview of the common techniques used for realistic quadruped animation.

1. Introduction

Human motion capturing has become a standard technique in computer graphic. In particular, the real looking animation of quadruped life is an area of active research. Virtual worlds, whether they be for interactive games or film, are made much more real and engaging by the inclusion of animals observable in daily life.

Quadruped animation is an interesting area in computer graphic. But it must address the laborious task of properly articulating four limbs during locomotion and movement. In fact, the primary challenge of animating a quadruped appears to be believable locomotion, e.g. coordinated foot movement adhering to well known gaits, minimal foot skate, and responsive traction.

To cover this gap, we adapt several well known techniques from computer animation to work with quadrupedal motion capture data, and report on according series of experiments in this work.

2. Similar works

In this section covers the closest related works for various areas.

2.1. Similar Retrieval works.

Efficient motion retrieval of motion capture data requires all data-driven methods. One of this called “Match Webs” to index motion capture databases are introduced in Kovar and Gleicher, 2004. This method has quadratic complexity in the size of the motion capture database, since a local distance matrix has to be computed comparing each pair of frames.

The same complexity holds for the computation of a neighbor graph structure [1]. It’s the same concepts of image preprocessing methodology.

Boolean features [5] are introduced but it only works with segment human motion capture data. [4] present a fast method to search for numerically similar poses and extends pose matching to motion matching by employing a so called lazy neighborhood graph.

2.2. Action Recognition.

In the field of motion recognition, a wide variety of techniques was developed, depending on different types of input signals. Using video sequences as input, employ temporal templates based on static vector images [2]. Here, the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence.

Schuld use local space-time features in combination with Support Vector Machine (SVM) classifier for action recognition (Schuldt, 2004).

Arikan use an interactively guided SVM to annotate entire motion capture database in an interactive process. Their approach works well on a small motion capture

database of American football motions used in their paper [1].

2.3. Motion Reconstruction from Video.

A lots of solutions has been investigated by computer vision community for 3D motion reconstruction like construction of statistical human pose models transforming 2D silhouettes and contours into 3D pose and motion [3].

Some prior knowledge from motion captured databases utilized for 3D reconstruction and physics based modeling for video based human motions (Wei and Chai, 2010).

In motion priors modeling, the prior knowledge from mocap database is sometimes embedded into implementation of some constraints. Employ geometric prior information about the movement pattern in markerless pose tracking process (Rosenhahn, 2008).

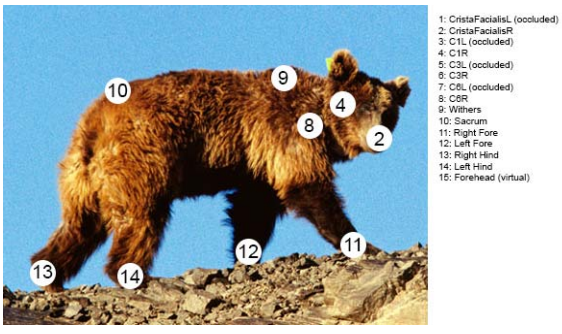
Most of the work regarding reconstruction from video sequences has been done on human motion like [9].

3. Quadruped Mocap Data

In this section, we present more details on the recording quadruped of our motion capture data. In this paper, we use three-dimensional kinematic data captured from mazaalai bear which lived in Mongolia.

3.1. Marker Setup

For recording, retro flective skin markers are attached to mazaalai using adhesive tape. Marker setups can be varied meeting the measurements for various purposes.



In a basic motion capture set-up of mazaalai, generally nine markers are required to capture the whole body motion. The first marker is normally placed on the head, then two on the trunk and the four on the hooves. However, the number of markers needs to be increased, when the measurement purpose is more complex and requires more details. In addition, marker setups can vary between subjects due to the size differences. In our case, markers are placed on the head (left and right crista facialis), on the highest point of the withers, sacrum and lateral side of each hoof to identify motion cycles.

4. Motion Retrieval

The search for similar motion point segments in a possibly annotated, database is a critical step in all data driven methods. We have decided to adapt the technique from Kr¨uger due to the following reasons: This technique can be easily parameterized with arbitrary feature sets.

4.1. kNN search

We separate between two types of kNN search: First, similar poses have to be found in a motion capture database. After computing feature sets for all frames of the motion database, the k nearest neighbors for a new pose can efficiently be retrieved by searching a kd-tree. Second, we search for the k most similar motion sequences compared to an example motion sequence. In this case, a technique called “lazy neighborhood” graph can be applied.

5. Action Recognition

We propose to use a modified k-nearest neighbor segments, that considers the temporal evolution of the regarded motion sequence. Instead of using the nearest neighbors obtained by a similarity search for voting directly, we consider poses for voting only, that are ending poses of a path through a lazy neighborhood graph (LNG) as described in Section 4.1. This lazy neighborhood graph can be parameterized with the width of the preceding window of frames. Thus, poses are regarded as similar, if the preceding window of frames is similar to the preceding query frames, too. Considering the temporal evolution of a motion segment, makes the knn voting more robust, due to the following reasons: First, poses that are numerically similar, but intersecting the actual sequence of poses from another direction will be filtered out. Second, if a query from a motion class is not reflected by the database and is given as input, k nearest neighbors can be returned from all motion classes. Thus, no

neighbors will be returned and the risk of wrong classifications decreases rapidly.

6. Conclusion and Future Work

In this work, we have used techniques for human motion data to quadrupeds. For motion retrieval, the basic techniques can be used without any modifications. Considering the results of the action recognition experiments, we show that extending the knn search by a temporal component, even a simple approach can lead to good results. The skeleton representation of quadruped might be computed and helpful in the process of full body quadruped motion reconstruction. The important step is the recording of other species in order to derive more general models of quadruped motion from such kind of data.

7. Reference

[1]. Arikan, O., Forsyth, D. A., and O'Brien, J. F., "Motion synthesis from annotations", ACM Trans. Graph., 2003, pp: 22:402–408.

[2]. Bobick, A. F., Davis, J. W., Society, I. C., and Society, I. C, "The recognition of human movement using temporal templates", 2001.

[3]. Elgammal, A. and su Lee, C., "Animal gaits from video", 2004.

[4]. Kr¨uger, B., Tautges, J., Weber, A., and Zinke, A., "Fast local and global similarity searches in large motion capture databases". In 2010 ACM SIGGRAPH/ Eurographics Symposium on Computer Animation, SCA '10, 2010, pp: 1–10,

[5]. M¨uller, M., R¨oder, T., Clausen, M., Eberhardt, B., Kr¨uger, B., and Weber, A., "Documentation mocap database hdm05". Technical Report CG-2007-2, Universit at Bonn, 2007.

[6]. Skrba, L., Reveret, L., Hetroy, F., Cani, M.-P., and O'Sullivan, C., "Quadruped animation", 2008.

[7]. Tautges, J., Zinke, A., Kr¨uger, B., Baumann, J., Weber, A., Helten, T., M¨uller, M., Seidel, H.-P., and Eberhardt, B., "Motion reconstruction using sparse accelerometer data", ACM Trans, Graph, 2011, 30(3):18:1–18:12.

[8]. V¨ogele, A., Hermann, M., Kr¨uger, B., and Klein, R., "Interactive steering of mesh animations", In 2012 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2012.

[9]. Yasin, H., Kr¨uger, B., and Weber, A, "Model based full body human motion reconstruction from video data". In 6th International Conference on Computer Vision, Computer Graphics Collaboration Techniques and Applications (MIRAGE 2013), 2013.

Installment For Measuring And Sharing Pm 2.5 Air Polution Concentration Through Social Media

Unursaikhan Batbayar[1], Sereeter Lodoysamba[1], Christa Hasenkopf[2], Joe Flasher[3]
[1]*School of Engineering and Apply Science,*
National University of Mongolia, Mongolia
{batbayar.unursaikhan, lodoysamba}*@gmail.com*
[2]*University of Colorado*
christa.hasenkopf @gmail.com
[3]*Ars Sollertia*
joseph.flasher @gmail.com

Abstract

In this paper, we describe hardware and software solution for measuring and announcing by Twitter and Facebook PM2.5 (particulate matter that is 2.5 microns in diamater or smaller) air pollution of Ulaanbaatar city (UB), the capital and largest city of Mongolia. Ulaanbaatar is one of most polluted cities of the world, with PM pollution concentration is exceeding more than 30 times of World Health Organization (WHO) standard. It is important to inform the public of air pollution levels by using social media such as Facebook and Twitter.

Measurement of PM2.5 concentration is made by DustTrack, TSI, and software was developed using LABVIEW to control collection of data real time every 30 min. It tweets and posts on Facebook every 3 hours, averaging the concentration.

The system consists of two main programs which are a control, a getting data from DustTrack(controlling program) and sharing information through social media (UB Data Platform). Reason of controlling DustTrack was there is no a real time communication program on DustTrack that we use. Controlling program uses USB virtual LAN of the instrument. There were some problems on local connection because of instrument's execution delay. We solved those problems with customized algorithms. Also we added some feature in control software for managing and scheduling the instrument. Diversity of the system is the controlling and sharing programs run in different countries. We used DropBox for exchanging data between the two programs.

The UB Data Platform is composed of a data ingest engine, a posting mechanism and a task scheduler. For the initial deployment, focusing only on UB Data, the data ingest engine calls out to Dropbox and checks for the existence of a data file that is shared publicly on the internet. If there is new data in this file, the values are stored in a database (http://www.mongodb.org/). After data storage, the new values are sent to pre-defined social channels, such as Facebook and Twitter via a public, documented API. The task scheduling is currently be handled by a cron job set to run every 3 hours and connecting to the Platform using a private key for authentication.

In addition to the read + post cycle outlined above, the platform also provides an un-athenticated API to get recent values (http://data.sciencerely.org/mongolia/api.html). The Platform is built using Node.js (http://nodejs.org/) to handle both data ingest, and posting. Finally, we were informing air pollution concentration one year round continuously from National University of Mongolia site of UB city. More than 6000 liked/followed our pages on Twitter and Facebook, more than 300 visitors see every twitting or posting. Additionally, this data from our pages have been shared by citizens with Mongolian Government leaders, attracted national and international media attention, and provided a way for the Ulaanbaatar public to critical access to air quality data. Work is currently being done to generalize the data ingest, posting mechanism and task scheduler to make the system work for a variety of instrument types and data, in addition to air quality data alone.

Results from PM2.5 measurements shared over social media by our instrument for the past year are presented in this paper.