行政院及所屬各機關因公出國人員報告書

（出國類別：會議）

參加「2014 資料探勘、網際網路運算及巨量資料國際會議」

服務機關：　行政院環境保護署

姓名職稱：　楊毅得 設計師

派赴國家：　馬來西亞

出國日期：　104 年 11 月 16 日至 11 月 20 日

報告日期：　105 年 3 月 18 日

目　錄

參加「2014 資料探勘、網際網路運算及巨量資料國際會議」出國報告

壹、 會議源起與目的

「2014 資料探勘、網際網路運算及巨量資料國際會議」係提供政府部門、非政府組織、研究機構及民間部門討論計算機工程、多媒體、人工智慧、資料探勘、網際網路運算、巨量資料及網路資訊安全以及其他相關主題，提供所有的人員共享並促進他們的知識的機會。

本(103)年度會議於 11 月 17 日至 19 日在馬來西亞亞太科技大學舉行（圖 1，議程及會議資料如附），除我國外，尚有阿爾及利亞、澳大利亞、厄瓜多爾、中國大陸、埃及、法國、德國、印度、伊朗、日本、利比亞、摩洛哥、阿曼、菲律賓、波蘭、俄羅斯聯邦、新加坡、斯里蘭卡、突尼斯、土耳其及美國等國參加。

貳、 會議經過

一、11 月 17 日研討會

本次會議於 11 月 17 日上午 8:30 至 9:30 各與會成員報到後，於 9:30 至 10：00 開幕（圖 2 及圖 3）儀式後分以下議程進行:

(一)第一場專題研究發表會

Cloud and Mobile security: Challenges and future research directions（雲端化及行動化的挑戰及未來研究方向）主要內容如下：

1. 由於雲端服務及行動設備/app 可存取及儲存可識別個人的敏感資訊，簡報中介紹對組織、個人的安全及隱私的威脅，並強調相關的安全、隱私的挑戰及減災戰略，以及概述一些潛在的研究議題。
2. 資訊系統的安全在今天已是一個熱門的話題，尤其許多焦點

都集中在資料保護的議題上，隨著雲端服務的成長及行動裝置的增加，傳統的安全方案已經變得不太夠用。因此，最大的挑戰是持續關注雲端服務和行動裝置所產生的安全漏洞，思考如何應對這些不斷更新且具有高度破壞力的惡意程式與威脅。

(三) 第二場專題研究發表會

DLP-Technologies: New Directions and Trends（DLP-技術：新方向及趨勢）（圖 4）主要內容如下：

1. 現今每個商業領域都可以看到全球資料洩漏(data leaks)的增長，因洩漏被竊的資料可以造成直接的財務損失，例如：客戶流失、金錢損失、聲譽損失以及負面印象。也可以造成間接損失 ，如資料傳輸到競爭對手所造成的影響。
2. 簡報中介紹什麼樣的信息可以得到保護，什麼樣的技術可以幫助我們保護這些資料，概述了 DLP-技術，並介紹未來的方向及趨勢。

(四) 參加「Computer Science, Computer Engineering, and Education Technologies」分組

1. 下午在不同會議室分 5 項主題進行，選擇其中「Computer Science, Computer Engineering, and Education Technologies」分組，包括 Layered Model Abstract Framework for Ubiquitous Learning Environment（無所不在學習的分層模型抽象框架）、Presenting New Method To Optimize Query In Distributed Database System（分散式資料庫最適化查詢的新方法）…等。
2. 其中印象較為深刻的是該研究指出無處不在的學習正在成為一種新的學習趨勢，無所不在的學習環境（ULE）是類似 OSI

的分層架構。因移動及手持設備倍數成長以及彈性的 IT 基礎設施、無線通信和都讓使用者更易於學習及取得知識。

二、11 月 18 日研討會

(一)第一場專題研究發表會

Using Fuzzy Logic to Evaluate Trust in E-Commerce（採用模糊邏輯評估電子商務信任）主要內容如下：

1. 信任被廣泛認為是以企業對消費者的電子商務模式(B2C EC)持續發展的一個重要因素。很多信任模型已經被開發，但多數是主觀的，在網路上進行交易，未考慮到電子商務的信任和客戶的直覺和經驗的模糊性和歧義。
2. 簡報中使用模糊推理來評價電子商務的信任模型的開發和實施。認為模糊邏輯是適合信任評價，因為它考慮到不確定性關係。

(三) 第二場專題研究發表會

Gamification of Teaching and Learning Activity: Prospect and Challenges of Mobile Game-based Learning（教學及學習活動遊戲化：手機遊戲學習的願景及挑戰）主要內容如下：

1. 行動裝置及 apps 的成長預示未來全球正規教育的學習將以遊戲為基礎，遊戲開發商及學者利用移動計算的好處，促進玩手機遊戲學習，然而，仍缺乏證明這可以保證實現預期的學習成果。
2. 我國臺灣大學參加由美國華頓商學院與全球大學評比機構 QS 合作，在費城舉辦的第一屆教學創新大賽（Reimagine Education），以全球首創的線上遊戲學習系統 PaGamO，得到了全球第一屆教學創新冠軍（Overall Winner），該系統將線上遊戲選角色、破關等形式，設計成學習機率的遊戲，

是本次簡報的最佳實踐。

(四) 參加「Data Mining, Internet Computing, and Big Data」分組

1. 下午在不同會議室分 5 項主題進行，選擇其中「Data Mining, Internet Computing, and Big Data」分組（圖 5），包括 Estimating Tea Stock Values Using Cluster Analysis（使用集群分析來評估茶的股票價值）、Server Monitoring Using Android Devices（使用 Android 裝置監控伺服器）、Feel the Heat: Emotion Detection in Arabic Social Media Content（阿拉伯語社群媒體內容情緒檢測）….等。
2. 其中印象較為深刻的是研究檢測阿拉伯語社交媒體網站（如 Facebook 和 Twitter）的情緒用字，其簡報說明作者表達在文字中的溝通風格明顯的會影響情緒的反應。

三、11 月 19 日參訪馬來西亞太子城(Tour to PUTRAJAYA) 相關建築並聽取解說

11 月 19 日上午是由馬來西亞亞太科技大學驅車前往馬來西亞太子城，參觀相關建築（如首相官邸，圖 6）並聽取解說。太子城，是馬來西亞政府建立的新市鎮，位於吉隆坡與吉隆坡國際機場之間。整個城市面積廣闊，也是馬來西亞計劃中的新行政首都。馬來西亞首相官邸及政府各部門陸續遷入辦公，住宅區、商業區、文化、休閒設施和交通體系也有基本配套措施。

參、心得與建議

一、 本次研究發表提及雲端服務、行動化(mobile)及電子商務資訊安全是未來需關注的挑戰及研究方向，包括資料洩漏、遺失、帳戶劫持、網路攻擊及惡意程式等，隨者越來越多的雲端服務及行動裝置導入，未來本署可思考導入相關資訊安全防護機制，以避免可能產生的問題。

二、 運用資料探勘(Data mining)技術，從巨量資料(Big Data)中取出隱含的過去未知的有價值的潛在資訊，本次發表會中有研究利用社群媒體、搜尋結果進行分析，從當中了解資料之間的相關性，以及預測未來發展趨勢。其應用皆可作為本署參考。

三、 持續參與相關國際會議，瞭解現階段環境資料探勘、網際網路運算及巨量資料之相關技術及未來發展。並參考各國相關研究及應用，作為我國未來導入參考。

圖 1 馬來西亞亞太科技大學


The Third World Congress on Computing
and
Information Technology
(WCIT 2014)

圖 2 開幕

圖 3 與會人員

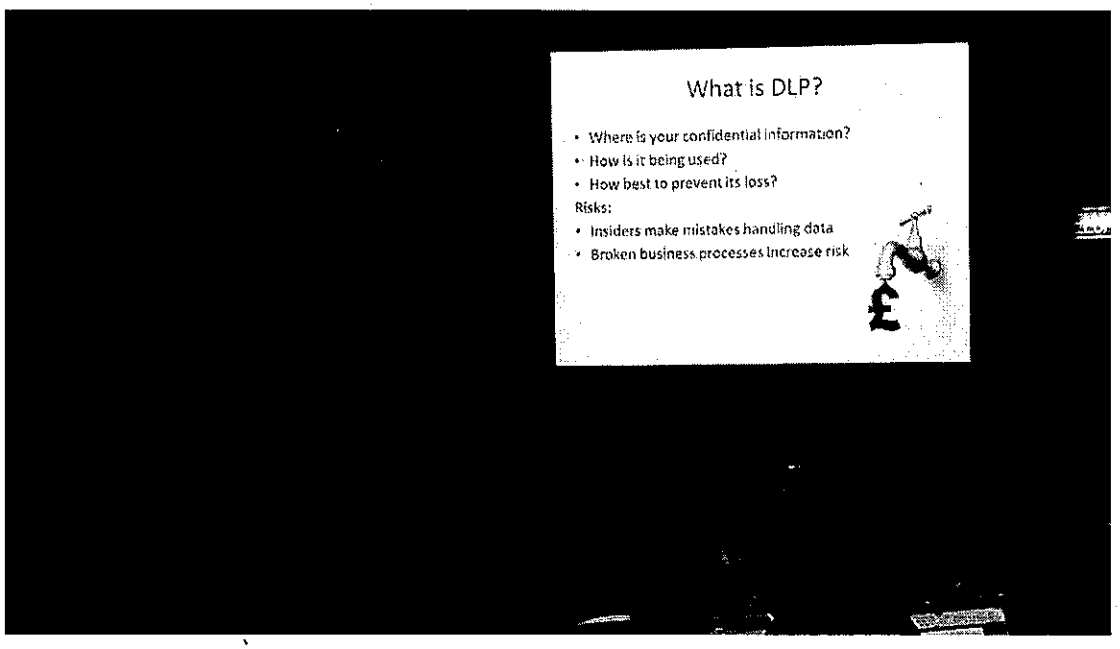


圖 4 DLP-Technologies: New Directions and Trends

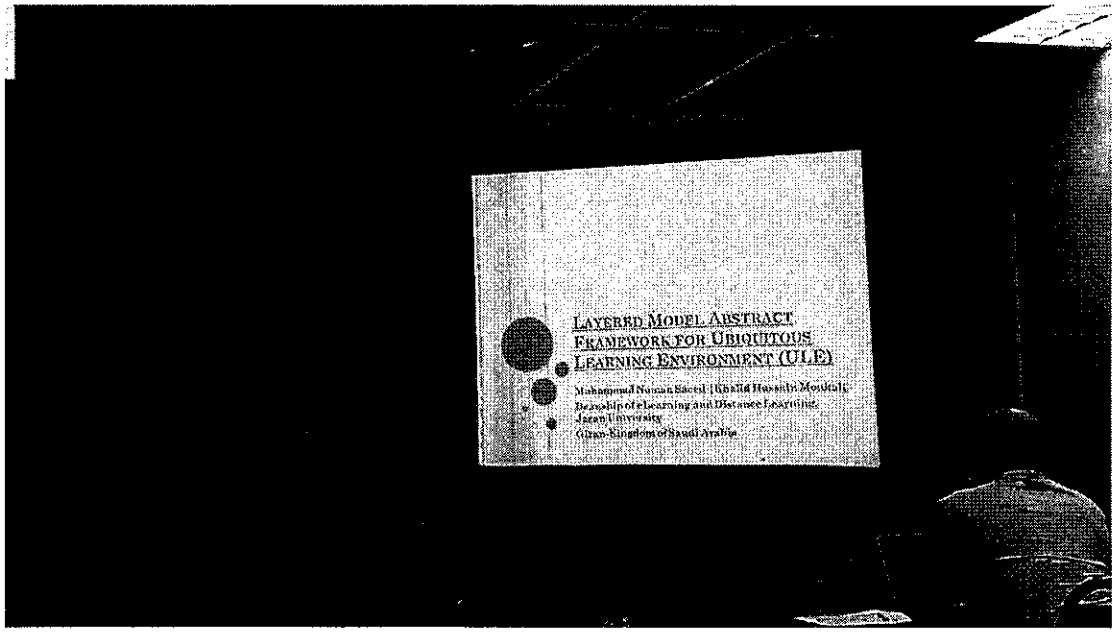


圖 5 Data Mining, Internet Computing, and Big Data

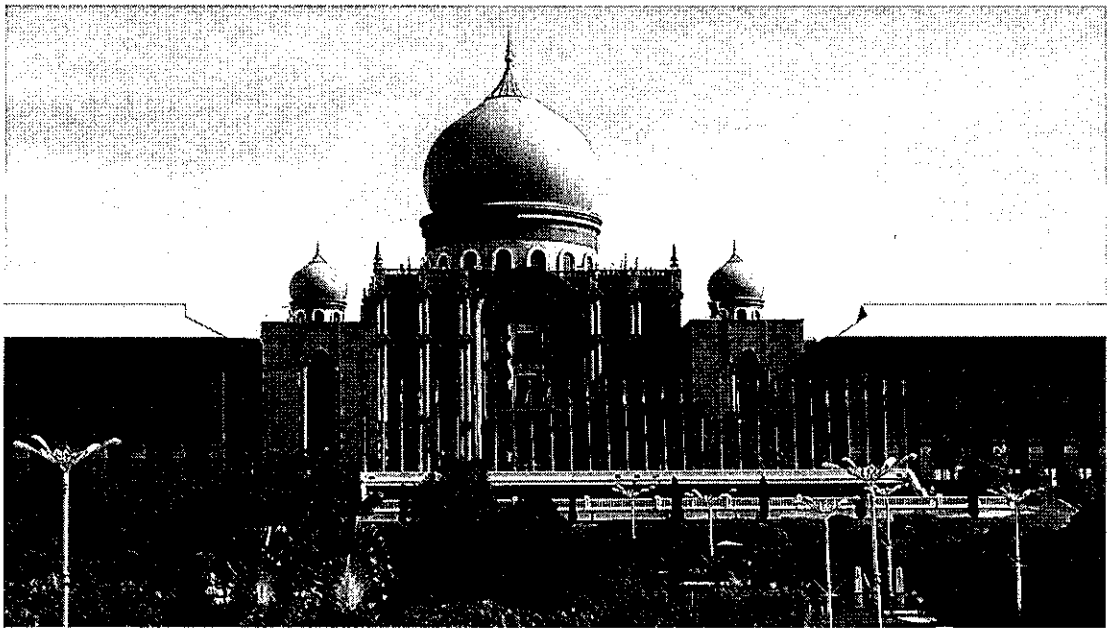


圖 6 參訪馬來西亞太子城（首相官邸）

附錄 會議資料

CONFERENCE PROGRAM

The Third World Congress on Computing and Information Technology (WCIT2014)

The International Conference on Computer Graphics, Multimedia and Image Processing (CGMIP2014)

The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014)

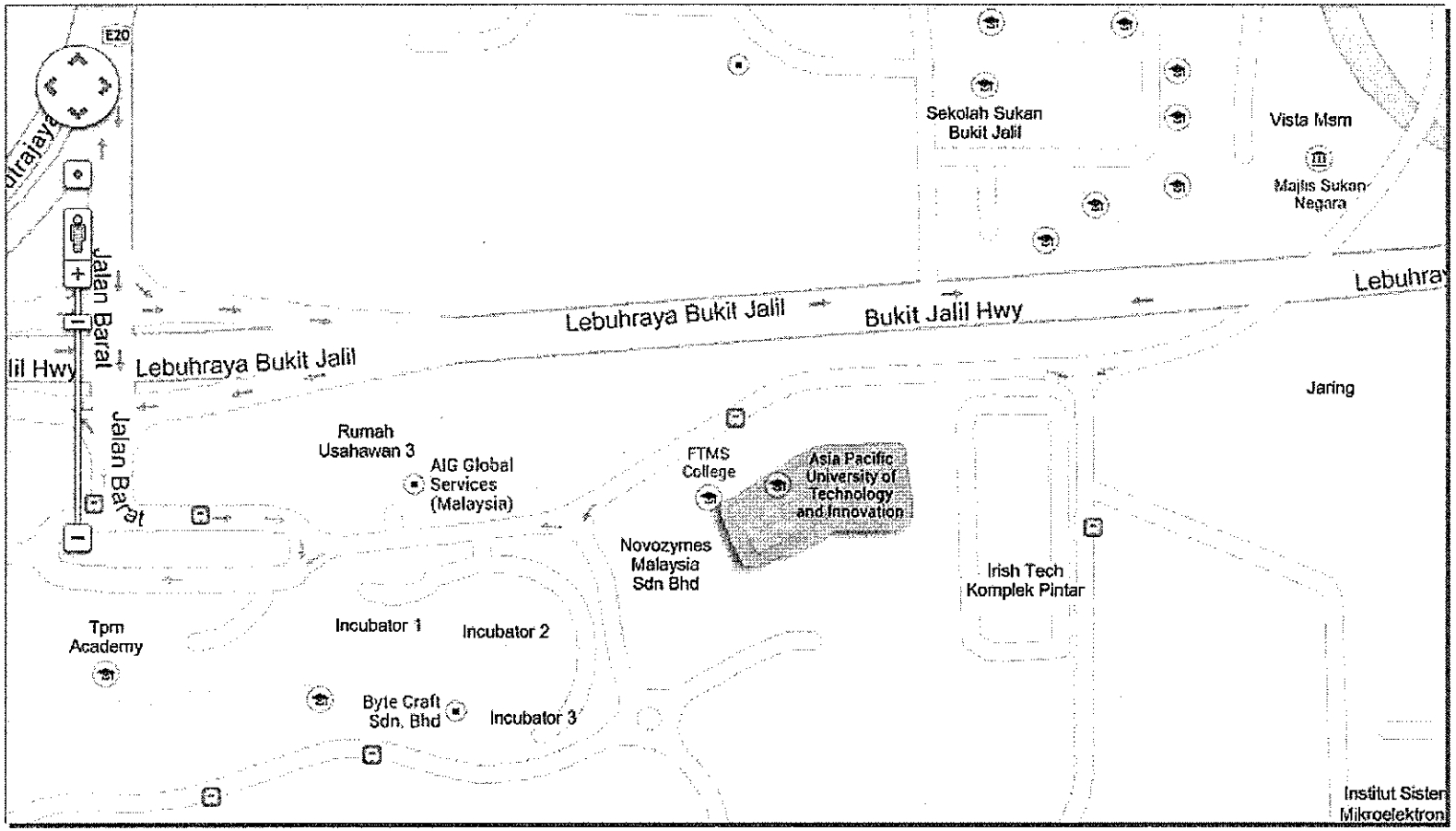
The International Conference on Computer Science, Computer Engineering, and Education Technologies (CSCEET2014)

The International Conference on Artificial Intelligence and Pattern Recognition (AIPR2014)

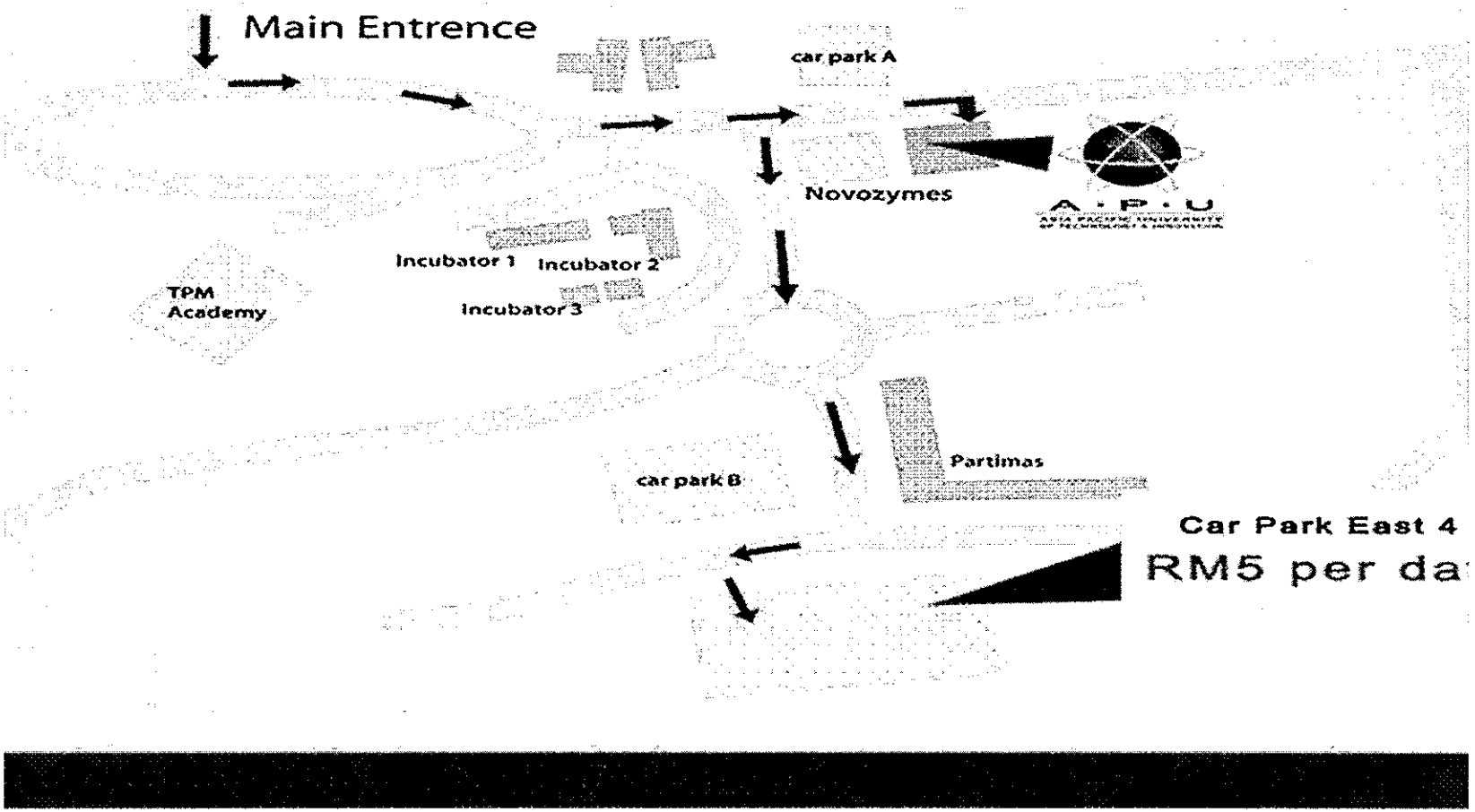
The International Conference on Cyber-Crime Investigation and Cyber Security (ICCICS2014)

The International Conference on Electrical, Electronics, Computer Engineering and their Applications (EECEA2014)

Notes:
The time of each presentation including questions is 20 minutes. Please adhere to it.
There are no presentations on Nov 19, only a tour.



Asia Pacific University of Technology and Innovation (APU)
Technology Park Malaysia
Bukit Jalil, Kuala Lumpur 57000, Malaysia
Tel: +603 8996 1000, 1 300 888 278 (Toll-free)
Fax: +603 8996 1001
Email: info@apu.edu.my, Web: www.apu.edu.my



Note: There are 5 parking lots has been reserved for special guests of the conference. It'll be labeled with 'SDIWC Conference' on the parking cone in front of the main building. For the participant parking lot, please refer to the map attached. As this parking slots are managed by TPM (Technology Park Malaysia), the parking fee will be RM5 per day.

Host Country: Malaysia

Participating Countries (sorted): Algeria, Australia, Brazil, Ecuador, Czech Republic, China, Egypt, France, Germany, Indonesia, Iran, Korea, Japan, Kuwait, Libya, Romania, Morocco, Oman, Philippines, Poland, Russian Federation, SA, Singapore, Sri Lanka, Thailand, Tunisia, Turkey, USA

WCIT2014 Schedule

Date	Time	Activity
Nov. 17, 2014 (Monday)	8:30 am – 9:30 am	Registration
	9:30 am –10:00 am	Opening Ceremony
	10:00 am – 11:00 am	Keynote Presentation by: *Raymond Choo* University of South Australia, Australia Keynote Title: **Cloud and Mobile security: Challenges and future research directions**
	11:00 am – 11:30 am	Tea Break
	11:30 am – 12:30 pm	Keynote Presentation by *Ekaterina Pshehotskaya and Tamara Sokolova* Keynote Title: **DLP-Technologies: New Directions and Trends**
	12:30 pm – 2:00 pm	Lunch Break
	2:00 pm – 5:00 pm	Session 1
	2:00 pm – 3:30 pm	Linguistic Workshop – Dr. Ekaterina and others
	3:30 pm	Tea Break
Nov. 18, 2014 (Tuesday)	9:30 am – 10:00 am	Briefing session of future events and introducing The Society of Digital Information and Wireless Communications (SDIWC)
	10:00 am – 11:00 am	Keynote Presentation: *Farid Meziane* University of Salford, United Kingdom Keynote Title: **Using Fuzzy Logic to Evaluate Trust in E-Commerce**
	11:00 am – 11:30 am	Tea Break
	11:30 am – 12:30 pm	Keynote Presentation by *Tan Wee Hoe* University Pendidikan Sultan Idris, Malaysia Keynote Title: **Gamification of Teaching and Learning Activity: Prospect and Challenges of Mobile Game-based Learning**
	12:30 pm – 2:00 pm	Lunch Break
	2:00 pm – 5:00 pm	Session 2
	3:30 pm	Tea Break
Nov. 19, 2014 (Wed.)	10:00 am – 1:00 pm	Tour to PUTRAJAYA

November 17, 2014 (Monday)

Title	Keynote Presentation
Venue	Auditorium 2
Time	10:00 am – 11:00 am
Speaker	Raymond Choo
Presentation Title	Cloud and mobile security: Challenges and future research directions

Title	Keynote Presentation
Venue	Auditorium 2
Time	11:30 am – 12:30 pm
Speaker	Ekaterina Pshehotskaya and Tamara Sokolova
Presentation Title	DLP-Technologies: New Directions and Trends

November 17, 2014 (Monday)

Title	Computer Graphics, Multimedia and Image Processing+ Simulation and Virtual Reality	Computer Science, Computer Engineering, and Education Technologies	Artificial Intelligence and Pattern Recognition	Electrical, Electronics, Computer Engineering and their Applications	WORKSHOP
Room	L1-2	L1-5	L1-7	L1-8	L1-9
Time	2:00 pm – 5:00 pm	2:00 pm – 5:00 pm	2:00 pm – 5:00 pm	2:00 pm – 5:00 pm	2:00 pm – 3:30 pm
Break	3:30 pm – 4:00 pm				
Chair	Dr Chen Tet Kuan	Dr Imran	Zailan Arabee	Dr Sathish Kumar	
Asst	Zety Marlia	Leong Swee Kee	Akansha	Vickneswari	
	121- Interactive Editing of Human Locomotion on an Arbitrary Motion Path (Yejin Kim- Korea) 125- Multilanguage querying for image retrieval (Arinori Takahashi- japan) 126- An Extended Framework for Visualizing the Data from Both Local Databases and Semantic Web Databases (Wei Shi- Japan) 132- Adaptive Video Watermarking Key Based on Multiband DWT & DCT & SVD (Hoda Farag- Eygpt) 131: A GPU based Real-Time Line Detector using a Cascaded 2D Line Space (Jochen Hunz- Germany) 128*- Semi-automated Cellular Tomogram Segmentation Workflow (CTSW): Towards an Automatic Target-scoring System (Nur Intan Raihana Ruhaiyem- Malaysia) 114- Surface Tension Approximation in Semi-Lagrangian Level Set Based Fluid Simulations for Computer Graphics (Israel Pineda- Ecuador)	108 - Layered Model Abstract Framework For Ubiquitous Learning Environment (Ule) (Muhammad Noman Saeed – Saudi Arabia) 130 - An Emotional User Interface Authoring Framework For Mobile E-Learning Applications (Kyu-Wan Kim - Korea) 128 - Shifting Virtual Reality Education To The Next Level – Experiencing Remote Laboratories Through Mixed Reality (Max Hoffmann - Germany) 126 - The Design Of Interactive Assessment-Cognitive Schema-Based System: An Exploratory Study In E-Learning Implementation (Melvin A. Ballera - Libya) 112 - Coordinating Mobile Servers For Static Hierarchical States (Savio S. H. Tse - Turkey) 141 - Towards Constructing A Platform That Makes Learning Contents On The Web "Anti-Ubiquitous" (Noriki Amano - Japan) 144 - Extracting Agent-Based Models For Considering Cultural Factors Using Multilingual Case Method System (Kenji Terui - Japan) 131 - Presenting New Method To Optimize Query In Distributed Database System (Sajjad Baghernezhad - Iran)	106- Quantum Cryptography Protocols with Tri-partite Entanglement (Hilal Al Hadhrami- Oman) 110- Variables selection for multiclass SVM using the multiclass radius margin bound (Fatima Zahra Aazi- Morocco) 120- Dictionary Learning Using EMD and Hilbert Transform for Sparse Modeling of Environmental Sounds (Bochra Bouchhima - Tunisisa) 124- Validation of XML Document Content Using Ontology (Shinji Norimatsu - Japan) 142- Restricted Boltzmann Machines for Modeling Businesses (Andreea Salinca - Romania) 135- Interactive versus Passive 2D Face Spoofing Detection (Mohamed Moustafa - Egypt) 139- Empirically Comparing Three Multi-Objective Optimization Approaches for the Automated Evolution of Snake-Like Modular Robots (Wei Shun Chee - Malaysia) 123- Line Detection by Centre and Width Estimation (Wei Hua Chen - Australia) 119- On the Enumeration of Frequent Patterns in Sequences (Zied Loukil - Tunisia)	119 - Detecting Equivalent Mutants Using Symbolic Computation (Hirohide Haga – Japan) 105 - Engineering And Optimization Of Mobile Network For Maximum Coverage (Abdelkrim Khireddine – Algeria) 117 - Design Of An Iec 61850 Based Communication System For Der Management (Taein Hwang - Korea) 113 - Push-Pull Class-E Power Amplifier With A Simple Load Network Using An Impedance Metched Transformer (Jinhee Kwon – Korea) 123 - Investigation Of In-Situ Doping Profile For N+Pn+ Bidirectional Switching Device Using Si1-Xgex/Si/Si1-Xgex Structure (Il Pyo Roh - Korea) 120 - The Basic Principles Of Capacitive Blood Pressure Measurement Method And Wireless Data Transmission (Vitaliy Petrov - Russia) 129 - Integrated Secure Vehicle Reservation And Parking Management System Using Gsm And Short Messaging Service (Sms) (Sardar Ali - Malaysia)	What I Really Do As Linguist in Data Leakage Prevention (DLP) by **Dr Ekaterina Pshehotskaya and Tamara Sokolova/** InfoWatch – Russia

November 18, 2014 (Tuesday)

Title	BRIEFING
Venue	Auditorium 2
Time	9:30 am – 10:00 am
Speaker	
Presentation Title	Introduction & Future Events of The Society of Digital Information and Wireless Communications (SDIWC)

Title	Keynote Presentation
Venue	Auditorium 2
Time	10:00 am – 11:00 am
Speaker	Farid Meziane
Presentation Title	Using Fuzzy Logic to Evaluate Trust in E-Commerce

Title	Keynote Presentation
Venue	Auditorium 2
Time	11:30 am – 12:30 pm
Speaker	Tan Wee Hoe
Presentation Title	Gamification of Teaching and Learning Activity: Prospect and Challenges of Mobile Game-based Learning

November 18, 2014 (Tuesday)

Title	Computer Graphics, Multimedia and Image Processing+ Simulation and virtual reality	Data Mining, Internet Computing, and Big Data	Computer Science, Computer Engineering, & Education Technologies Electrical, Electronics, Computer Engineering & their Applications	Artificial Intelligence and Pattern Recognition	Cyber-Crime Investigation and Cyber Security
Room	L1-2	L1-5	L1-9	L1-7	L1-8
Time	2:00 pm – 5:00 pm	2:00 pm – 5:00 pm	2:00 pm – 5:00 pm	2:00 pm – 5:00 pm	2:00 pm – 5:00 pm
Break	3:30 pm – 4:00 pm				
Chair	Dr Mohd. Ahmadi	Wong Bee Suan	Dr Thomas	Dr Vazeerudeen	Dr Arkadiusz Lach
Asst	Hamzah	Rizawati	Yvette	Lee Kim Keong	Reza Adinehnia
	139- Development for 3D Video Communication System by Using Kinect and Head Mount Display in the AR Space (Hideyuki Hashimoto- Japan) 140-A Robust Recognition Method for Occlusion of Mini Tomatoes based on Hue Information and Shape of Edge (Fumiya Iwasaki- Japan) 134- Applying Augmented Reality Technology to Promote Traditional Thai Folk Musical Instruments on Postcards (Suwichai Phunsa- Thailand) 136- Car Plate Detection Engine Based on Conventional Edge Detection Technique (Hamam Mokayed- Malaysia) 141- Design and Implementation of Automatic Aerial Mapping System Using Unmanned Aerial Vehicle Imagery (M.S. Javadi- Malaysia) * Any re-scheduled presentation	111 - Comparison of Machine Learning Algorithms Based on Filipino-Vietnamese Speeches (Hoa Le - Philippines) 119 - Estimating Tea Stock Values Using Cluster Analysis (Amitha Caldera - Sri Lanka) 136 - A Method for Evaluating an Action Rule Specified by a User (Seunghyun Im - United States) 123 - Feel the Heat: Emotion Detection in Arabic Social Media Content (Omneya Rabie - Egypt) 132 - Towards Applying Support Vector Machine Algorithm in Employee Achievement Classification (Hamidah Jantan - Malaysia) 133 - Construction of subject-independent brain decoders for human fMRI with deep learning (Sotetsu Koyamada - Japan) 138 - Extraction of Automatic Search Result Records Using Content Density Algorithm Based on Node Similarity (Yasar Gozudeli - Turkey) 134- Server Monitoring Using Android Devices (Negar Shakeribehbahani – Malaysia	133 - Semiautomatic Porting Of The C Library (Ludek Dolihal - Czech Republic) 145 - The Quality Analysis Of The Video Game Failure (Tomoyasu Tanaka - Japan) 116 - Online Project And Assignment Submission, Management And Progress Monitoring System (Opas) (Poorya Bagheri Faez – Malaysia) 148 - Rapid Method For Embedded Systems Hardware And Software Education (Naohiko Shimizu - Japan) 149 - Implementation Of Game Tree Search Method By Using Nsl (Naohiko Shimizu - Japan) 125* - Cost Implication Analysis Of Ncomputing Adoption-A Case Study Of Tanzania Education Sector (Renatus Michael – Tanzania) 136* - Acceptance Of E-Marketing Strategies In Developing Countries- A Case Study Of Tanzania Smes (Renatus Michael – Tanzania) * Any re-scheduled presentation	118- Content Based Video Quality Control for Wide-area Video Surveillance Systems (Takeshi Arikuma - Singapore) 125- Policy Gradient Method Using Fuzzy Controller in Policies and Its Application (Noor Imanina Binti Noor Hasnan - Japan) 131- Text Classification Using Computational Model of the Cerebral Cortex (Koki Hatano - Japan) 116- Comparison of Classifiers for Retinal Pathology Images using SURF and Bag-of-Words Model (Fanji Ari Mukti - Malaysia) 122- Using Latent Semantic Analysis to Identify Quality in Use Indicators from User Reviews (Wendy Wei Syn Tan - Malaysia) 137- Integrating Evolutionary Robotics with 3D Printing for Rapid Fabrication and Deployment of a Physically-Simulated Autonomous Six Articulated-Wheeled Robot (shunhoe lim - Malaysia) 145- Best-Parameterized Sigmoid ELM for Benign and Malignant Breast Cancer Detection (Chandra Utomo - Indonesia) 143- Noisy Text Normalization Using an Enhanced Language Model (Mohammad Arshi Saloot - Malaysia) 127- Resource Acceleration of Autonomous Stochastic Moving Multi-Agents with Boundary Effects (Isamu Shioya - Japan)	108 - Analysis Of Slow Read Dos Attack And Countermeasures (Junhan Park - Korea) 119 - A Protection Architecture For Malicious Javascripts On Web Browsers (Woung Jang - Korea) 120 - Analysis of ISO27001 Implementation for Enterprises and SMEs in Indonesia (Candiwan Candiwan - Indonesia) 104*- Security Issue on Cloned TrueCrypt Containers and Backup Headers (Rodrigo Ruiz - Brazil) 127- A Review: Network Forensic Analysis Framework in IaaS Cloud Computing Environment(Samsiah Ahmad - Malaysia) 123 - WebKure: A Web Vulnerability Auditor (Rana Jacob Jose - Oman) 109 - The problems of investigation of identity theft in SNS (Arkadiusz Lach - Poland) 126* - Security in Depth Requires Secure Programming Languages Too (Walid Al-Ahmad – Kuwait) 102 - Cyber Crimes In Iran: Definition And Analysis (Mohsen Ghasemi Ariani - Iran)

		113* - An Approach to Detect Spam Emails by Using Majority Voting (Roohi Hussain - Pakistan)	126 - A Unique Signature Scheme Based On Candidate Multilinear Maps (Han Wang – China) 132* - Gpu-Based First Collision Detection In Parton Cascade In Heavy-Ion Collisions (Qing-Jun Liu - China) 122 - Anfis Based Intelligent Solar Flare Prediction System (Ajmal Hussain Shah - Malaysia) 125 - Real-Time Detection Of Suspicious Human Movement (Ka Fei Thang – Malaysia) 127 - The Potential Of Mobile Technology Application In Hypermarket Industry. A Case Of Malaysian Consumer Behavior (Salmiah Amin - Malaysia) 131 - Towards Green Computing Application For Measuring The Sustainability Of Data Centers: An Analytical Survey (Salmiah Amin – Malaysia) 134 - A Low Cost Vehicle Monitoring System For Fixed Routes Using Global Positioning System (Gps) (Chandrasekharan Nataraj - Malaysia) 133 - Design Of Gesture Technology Implementing Wireless Sensor Network And Short Message Service (Sudarmawan Sudarmawan - Indonesia)	133- Local Clustering Organization Based Subtour Extract Approach for Large-Scale TSP (Yohko Konno - Japan) 132*- Inference Engine for The Classification of Expert Systems Using Keyword Extraction Technique (Nabila Perveen – Pakistan 140*- Terrorist Group Prediction Using Data Classification (Faryal Gohar - Pakistan) 144*- Dynamic Bayesian Networks for Multi-Dialect Arabic Isolated Words (Elyes Zarrouk - Tunisia) 129*- Predicting Movie Incomes Using Search Engine Query Data (Chanseung Lee - USA) 126*- ComboSplit: Combining Various Splitting Criteria for Building a Single Decision Tree- Md Nasim Adnan - Australia)	

Layered Model Abstract Framework for Ubiquitous Learning Environment (ULE)

Muhammad Noman Saeed, Khalid Hussain Moukali
Deanship of eLearning and Distance Learning, Jazan University, Gizan-Kingdom of Saudi Arabia
msaeed@jazanu.edu.sa, kmoukali@jazanu.edu.sa

ABSTRACT
In last couple of years of Learning and Technology, we have observed the exponential growth of ICT tools, especially Mobile and other handheld devices along with the established and resilient IT infrastructure. This is already strengthening the IT services including Wireless Communication and availability of several other devices for user easiness and enables new epoch for learning and knowledge, especially for smart cities which promises to have a modern communication infrastructure. The development transforms the traditional learning to context responsive learning as well. Ubiquitous Learning is becoming a new learning trend, based on Learner centred approach. The research work in this paper will present an overview of the Ubiquitous Computing, Ubiquitous Learning Environment (ULE) similar to the OSI layered architecture.

KEY WORDS
Ubiquitous Learning, Ubiquitous Computing, Context Awareness, Wireless Communication, OSI Layer

1. INTRODUCTION
E-Learning can be defined as an educational process and scheme in which the substantial portion of teaching is conceded out through online distribution of knowledge for education & learning using Information and Communication Technologies and tools. It provides a new paradigm for learning pedagogy by using a set of IT enabled tools in classroom learning and using interactive assessment and lecture delivery with the help of audio, video, textbook and other material sharing as compared to the traditional learning and teaching environment.

The recent rapid developments of wireless and sensor technology in metropolis, the traditional network learning for the acquisition of knowledge, i.e. Transformation of knowledge from Electronic to Mobile Learning and then further extend to unconventional networked learning i.e. Context aware Learning or Ubiquitous Learning (U-learning). This remarkable change is due to the exponential growth of ICT infrastructure and strengthening the Wireless sensor network (WSN). The major difference between these ICT based learning pedagogies with the traditional methods is the relationship of the Learner in each scenario. E.g. the influence of learning experience of individual learner in a Ubiquitous learning environment is much higher than the presence at a Distance or Electronic learning. Also the learning was transformed from Individual experience learning to collaborative and shareable learning experience which increases from electronic to mobile and finally in Ubiquitous learning. The implementation of such ICT technologies using Wireless and Mobile communication results a Virtual Wireless University however there are several observers who assert that such implementation is firstly for financial and technological reflection rather than pedagogical. Nevertheless it is evident that the ramification for the integration of ICT specially the Wireless and Mobile communication enhances the Ubiquity, Sophistication, Compatibility, Emphasis, Savings and Standardisation in teaching and learning environment [14].

The recent development helped a learner for accessing the global communication and educational resources with its own pace and easiness. Such behaviour of ubiquitous computing nature with blending Ubiquitous Learning creates a new step in learning. The overall integration of ICT tool for providing omnipresence learning creates a huge change in the delivery of Education knowledge and sharing learning experiences.

ISBN: 978-1-941968-02-4

2. UBIQUITOUS COMPUTING & UBIQUITOUS LEARNING

In this section we will outline information for Ubiquitous Computing and Ubiquitous Learning with its evolution from eLearning, mLearning.

2.1 Ubiquitous Computing (UC)

There are many diverse definitions and ideas for Ubiquitous Computing; however Mark Weiser, who is considered as a father of Ubiquitous Computing named this technology as the Third Wave of Computing that resides from 2005-2020. First was the Mainframe systems that were used in shared mode and the users are using its high power processing typically for large calculation. After the era of Large Mainframes we have utilized Small Personal Computing age, in which a user has its own environment for accessing his personal computing. Finally, he determined a new era which he named as the Age of Calm Technology i.e. Ubiquitous Computing where technology recedes into the background of our lives. [1].

The other view about Ubiquitous Computing was presented by Gabriel, who stated that the Ubiquitous Computing era encompasses a series of computers or computing devices that are available for use as a sharing node for us. These computers cab are the devices that we may use for browsing the Internet and the others will be implanted in the walls, chairs, clothing, light switches, cars. In short Ubiquitous Computing is primarily considered an integrated world of computers and computing devices from large scale to microscopic [2].Today, we can name such type of small server network as collaborative microscopic or wireless sensor network.

The recent researchers may further clarify the Ubiquitous Computing term as a new hype in ICT world. Like previously mentioned, Ubiquitous Computing considered as a huge collection of integrated and tiny electronic computing devices (small computers) which have ability for communicating and computation. We can find such devices as smart mobile phones, contactless smart cards, handheld terminals, bar codes, sensor network nodes, Radio Frequency Identification (RFIDs) etc. Infact Ubiquitous Computing and such devices are now in our life and everywhere around us [3].

2.2 Ubiquitous Learning

Ubiquitous learning is reported to be both determined and omnipresent and allow access for learners to obtain immaculate knowledge and learning [13]. In line for its omnipresence nature, Ubiquitous learning has a possibility to transfigure the existing traditional as well as E-Learning and remove any physical constraints. Ubiquitous learning changes the Electronic Learning paradigm by proposing new learner-centred approach which can easily integrate within our education live. U-Learning is considered as a prominent swing of learning paradigm from traditional learning to e-Learning and m-Learning [See Figure-1,2].

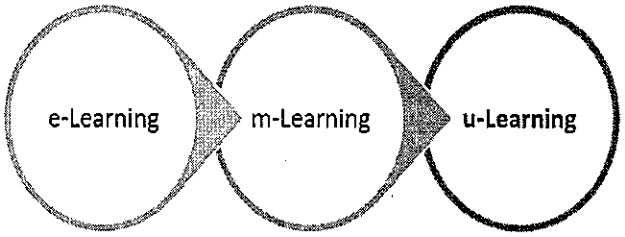


Figure-1 (Evolution of eLearning)

Based on various research on studying of the above evolution of ubiquitous learning, u-learning can easily be formulated as below from the amalgamation of Mobile Learning into its pre successor i.e. Electronic Learning in order to generate a purely ever-present u-learning environments [5].

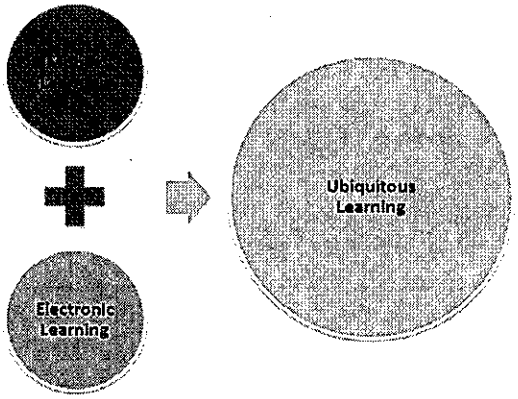


Figure-2 (Integration of Learning technologies)

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Additionally, the assimilation of adaptive learning (an individualized method of teaching) both with Ubiquitous Computing Learning produces extraordinary improvement for dissemination of Knowledge and learning on the basis of student/learner context and easiness.

3. UBIQUITOUS LEARNING FEATURES

Ubiquitous Learning is still considered as a new form for research and discussion nevertheless there is still confusion for highlighting its complete topographies. However, at this point of discussion and Based on the several studies carried out by the researcher[8,9,10,11,12] we can conclude that the studies for these investigators are somewhere overlapped with each other and therefore we can conclude to present the following main features model for u-learning.

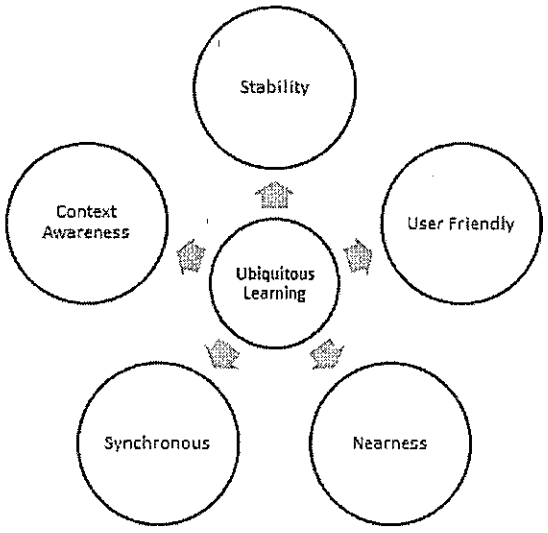


Figure-3 (Ubiquitous Learning Features)

The above cited features can be demonstrated as below

- Stability
  In U-Learning, the required and accurate information should always be available for the Learner at all time. The Stability feature of U-Learning determines that such information easily transmitted to the user / learner of the system unless the Learner deliberately eliminate or update it from the system.
- User Friendly
  The learning scenario in U-Learning must be available to Learner as per his requested time in easy and user friendless manner.
- Nearness
  At U-Learning a Learner must have the requested information, regardless of the location it stored.
- Synchronous
  This is one of the main features of the U-Learning system. As the whole environment is used for collaborative learning therefore a Learner in this system must be able to receive the information synchronously from its instructor, co-student etc.
- *Context Awareness*
  Last but not the least, context or location awareness in the ubiquitous learning is a must feature for a system to provide accurate requested information based on its respective environment/presence.

4. UBIQUITOUS LEARNING ENVIRONMENT (ULE)

Ubiquitous Learning standpoints are based on Ubiquitous Computing Technology and Environment. Therefore U-Learning environment integrates several components from physical to intellectual including Learner location at Home, School or in Society etc. [6]. Such integration promises that a system user/learner can receive the information at anywhere, anytime, or any device. Another version of ULE that was presented to state that Ubiquitous environment is a situation created for the learner in which he can become totally engrossed in the whole learning phase which is happening around him. Diagrammatically the ULE can be seen as Figure-4. [12]

ISBN: 978-1-941968-02-4 ©2014 SDIWC

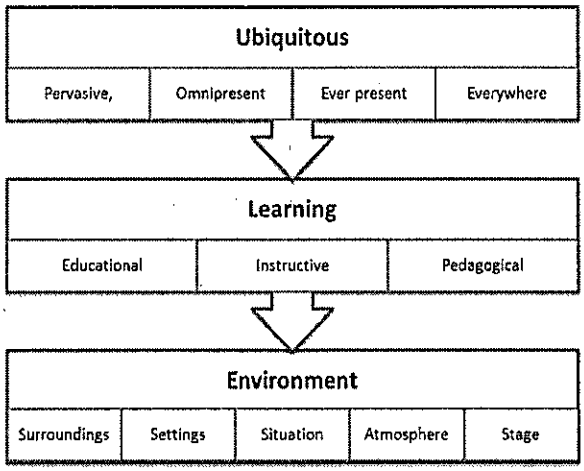


Figure-4 (Ubiquitous Learning Environment)

Based on the above defined research, a Ubiquitous Learning Environment is a framework for spreading context aware learning to its user which is particularly referred to be as a Learner. ULE is an integrated learning environment, which is cohesive from various other factors like physical, social, informational and technical settings [6]. The basic intention of developing the ULE is to integrate the different entities varies from the institutional areas, personal links, societies etc. effortlessly.

5. LAYERED MODEL FRAMEWORK UBIQUITOUS LEARNING ENVIRONMENT

Mobile technologies have many advantages—they are ubiquitous, portable, and easy to use and can deliver audio, video, multimedia, and text—and the abundance of educational applications developed for these platforms makes them a highly promising mode of teacher professional development [7]. The modern development of Mobile infrastructure has shattered as tools for communication as well as a tool for learning and teaching by students, teachers, etc. This type of Mobile Learning which is happening everywhere due to excellent Wireless or Mobile communication coverage gradually turned and termed as Ubiquitous Learning and urged researchers to design efficient ubiquitous learning environment based on these technologies and equipment. Based on the basic E-Learning Three (3) tier Elements, i.e. Content Delivery, Authoring Tools & Learning Management System [4], U-Learning can also present in the same hierarchical way as highlighted in Figure-5.

Due to the latest advancement in wireless and mobile communication, there are many researchers who are working to create a ubiquitous learning environment from a small setting like Smart Classroom to large setting like U-Learning in Smart Cities. We also endeavour to develop a Ubiquitous Learning Environment based on the OSI (Open System Interconnection) layer model. The Open Systems Interconnection (OSI) model is an abstract layer model that presents and regulates the internal functions of any communication system by segregating it into virtual layers. The idea is to mix the communication model with a learning model for better understanding the communication and learning scenarios. Below is our proposed model which demonstrates the layered abstract framework for designing Ubiquitous Learning Environment.

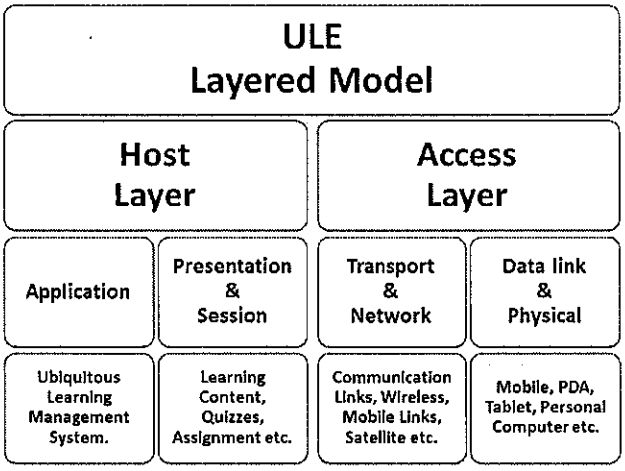


Figure-5 (ULE Layered Model)

As shown in above figure, the hierarchical based layered model is explained below.

- Host Layer
  This proposed layer for ULE model consist of two sub layer which holds the information of the ubiquitous learning system and learning content. The Host layer of ULE environment is the one which is closest to the Learner and Teacher which means that this layer is directly interacting with an environment which will be used for

managing and utilizing the ULE system. The Host layer is further classified in two (02) sub layers as based on the OSI model.

- Application Layer
  This sub layer of Host layer represents the ubiquitous learning core system, i.e. Ubiquitous Learning Management System (ULMS). This layer has a core responsibility for smooth functioning of ULMS to provide services to Learner and other user of the system as per their demand.
- Presentation & Session Layer
  This sub layer of Host layer presents the Learning content for Learner through ULMS. This layer is also providing session stability while conducting Assessment through Quizzes, Assignment etc. This hybrid layer of ULE model provides the session orientation between the Learner and the Assessment application that are running under ULMS at Application layer.

- Access Layer

The second core layer of the proposed ULE model is an Access Layer. This layer of ULE model provides and responsible for the Accessibility of Learner to the Host layer for gain access to ULMS and learning material. This layer also consist the below two sub layer consist of network connectivity and hardware for retrieving ubiquitous information from ULMS. The Access layer is responsible for calibrating the various Hardware devices that will be used for accessing and providing Ubiquitous learning.

- Transport & Network
  This layer contains communication network and responsible for efficient wireless and mobile connectivity. Also the layer control the wireless sensor, RFID, QR code, etc. communication as well if use by hybrid ubiquitous environment. This link is crucial for whole learning process because it provides and ensure the network connectivity and communication link availability between the devices and application for ubiquitous learning.
- Datalink & Physical
  U-Learning fundamentally comprises learning through handheld devices as well as networked devices. This layer is correspondent to the OSI low level layers which covers the hardware like Tablets, Personal Digital Assistant, Media/Audio devices, Smart Phones etc. and for accessing the learning content based on the methodology of anytime, everywhere learning.

6. CONCLUSION AND FUTURE WORK

In traditional eLearning, distance learning or even mobile learning, the learner can attain knowledge by use of predefined knowledge which is dedicated to them by the trainer. However, in Ubiquitous learning environment the Learning Management System is such designed which help learner to learn and gain knowledge based on his willingness and Context aware environment. In this paper, a context-aware approach for learning, i.e. Ubiquitous Learning with an abstract model is proposed to design and understand a smooth ubiquitous learning environment. Also, we presented the design model of ULE (Ubiquitous Learning Environment) based on OSI (Open System Interconnection) layer model. We have described the Ubiquitous Computing, Ubiquitous Learning and its features and lastly presented our model. The proposed model can be helpful for developing an actual ULE learning infrastructure for providing anytime learning. The new revolution of learning environment due to the U-Learning and E-Learning ensure the future growth of Ubiquitous & Mobile technology for the new learning environment. Due to the rapid growth of such learning, the future areas of

ISBN: 978-1-941968-02-4 ©2014 SDIWC

research under this area also covers the security issues that can arise in Information Technology (IT) side of ULE learning and needs a secured Learning environment for IT engineers.

7. REFERENCES

[1] M. Weiser, "The Computer for the Twenty-First-Century, Scientific American Special Issue on Communications, Computers, and Networks"(1991).

[2] M. Weiser & J.S. Brown, "The Coming Age of Calm Technology"(1996).

[3] K. Sakamura, N. Koshizuka, "Ubiquitous Computing Technologies for Ubiquitous Learning,", IEEE International Workshop on Wireless and Mobile Technologies in Education (2005) pp.11-20.

[4] O.K. Boyinbode, K.G. Akintola, "Effecting E-Learning with U-Learning Technology in Nigerian Educational System", The Pacific Journal of Science and Technology (2009), Volume 10. Number 1, pp. 204-210.

[5] D. Casey, "U-Learning = E-Learning + M-Learning", Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (2005), Vol 1, pp. 2864–2871.

[6] L. Li, Y. Zheng, H. Ogata, and Y. Yano, "A Conceptual Framework of Computer-Supported Ubiquitous Learning Environment", Proceedings Of The IASTED International Conference On Web Based Education, (2005).

[7] S. Pasnik, "IPod in Education: E-Potential for Teaching and Learning - A White Paper for IPods in Education", Prepared For Apple Computers. New York (2007)

[8] Y. Shyan, T. Chien, J. Ping, C. Yu, "A Mobile Scaffolding-Aid-Based Bird Watching Learning System", IEEE International Workshop on Wireless and Mobile Technologies in Education. (2002). pp 15-22.

[9] M. Curtis, K. Luchini, W. Bobrowsky, C. Quintana, E. Soloway, . "Handheld Use in K-12: A Descriptive Account" IEEE International Workshop on Wireless and Mobile Technologies in Education. (2002). pp. 23-30.

[10] H. Ogata, R. Akamatsu, Y. Yano, "Computer Supported Ubiquitous Learning Environment For Vocabulary Learning Using RFID Tags", Technology Enhanced Learning IFIP International Federation for Information Processing (2005) Volume 171, pp 121-130

[11] G. J. Hwang, C. C Tsai, S. J. Yang, "Criteria, Strategies and Research Issues of Context-Aware Ubiquitous Learning", Educational Technology & Society (2008), Vol.11, No.2, pp.81-91.

[12] P.S.Chiu, Y.H. Kuo, Y.M. Huang, T.S. Chen, "A Meaningful Learning Based U-Learning Evaluation Model", Eighth IEEE International Conference On Advanced Learning Technologies, (2008) pp. 77 – 81.

[13] V. Jones, J. H. Jo, "Ubiquitous learning environment: An adaptive teaching system using ubiquitous technology. Australasian Society for Computers in Learning in Tertiary Education (ASCILITE), 2004 proceedings.

[14] M. Thomas, "E-Learning on the move". The Guardian May (2005).
http://tiny.cc/7mm1ox

An Emotional User Interface Authoring Framework for Mobile E-Learning Applications

Eun-jung Lee*, Kyu-wan Kim*, Woo-bin Kim*, Byung-soo Kim** and Mi-Ae Kang**
Kyonggi Univ*, Yaginstek Co.,Ltd**
ejlee@kyonggi.ac.kr, kkw5240@naver.com, woobinkim508@naver.com, bskim@yagins.com, makang@yagins.com

ABSTRACT

The development of mobile applications is an emerging issue in research on e-learning. One of the most difficult problems in particular is developing qualitative user interfaces (UIs). Here we consider a UI design framework that facilitates the development of rich emotional user interactions and integration of design results with native codes. We present the three-tier architecture of an authoring framework for developing platform-independent UIs with visual effects and animations. We also describe the implementation of a prototype, the DAT4UX system, which demonstrates the feasibility of our design proposals.

KEYWORDS

Mobile E-learning, Mobile App, User Interface, Mobile Application Development, Emotional Awareness.

1 INTRODUCTION

Developing mobile applications for e-learning software is a challenge because of the multiplicity of platforms, lack of skilled programmers, and rapid change of technologies. Although many approaches provide frameworks for developing cross-platform applications, they are more programmer-oriented and based on specific technologies than are web development tools[1], [2].

Similar to the early days of web technology, collaboration between designers and programmers is one of the main challenges[3], [4]. For designers working on mobile platforms, the difficulties stem from 1) coping with platform-dependent image rendering mechanisms and various screen sizes, 2) integrating the design result with the program codes, and 3) lack of tools to create animated and dynamic interactions. From the viewpoint of management, the variety of platforms and the lack of helpful tools also raise the cost of developing mobile applications.

To mitigate these problems and smoothly coordinate the collaboration between designers and developers, we need an authoring tool that allows designers to create platform-independent user interface (UI) designs on their own and integrate them into the program codes with minimal programming efforts. After inspecting various alternative architectures, in this paper we present an approach for designing visual authoring tools for mobile multi-platform applications, and implement it in a prototype called DAT4UX. Our contributions are four-fold:

1) Several architectural alternatives are considered for mobile e-learning applications. In particular, we discuss the roles of the native application component in various hybrid application structures.

2) After reviewing current frameworks for platform-independent UI authoring, we describe a three-tier architecture that consists of a front-end, an intermediate representation (IR), and a back-end. The back-end manages most of the platform-dependent issues so that the front-end is free from such problems.

3) We describe an approach for designing authoring frameworks for dynamic UI behaviors in terms of animations and event handling.

ISBN: 978-1-941968-02-4

4) We describe the implementation of a prototype of the proposed model.

This paper consists of the following sections. In the next section (Section 2), previous work on the development of mobile e-learning applications is discussed. Subsequently, Section 3 gives an overview of the architectural alternatives of mobile e-learning applications. The design of the platform-independent authoring framework is given in Section 4. Implementation of the prototype DAT4UX system is presented in Section 5. We conclude by discussing our experience..

2 BACKGROUND

Interface design is a critical aspect of e-learning systems, as it has a significant influence on learning effectiveness [5], [6]. Although the design of the information structure and user interface should be based on educational models and learning activities, building a qualitative and reactive user interface remains a practical goal of learning system development.

Recently, many studies have investigated the application of mobile learning models to various areas [7], [8]. However, few studies have focused on the development of mobile learning applications and user interface design. Although several studies have developed and applied mobile learning clients to feature phones [9], [10], they need to be extended to smart mobile devices which are in widespread use at present.

The development of multi-platform mobile applications remains an active research topic. There are two main trends: web and native [11], [12]. Moreover, as discussed in several recent surveys [2], [3], [12], there remain many challenges to mobile app development. In order to overcome the variation in different computing environments, platform-independent frameworks for designing and developing multi-platform mobile applications have been created. Some of these frameworks are based on web technologies such as HTML, JavaScript, and jQuery [13].

On the other hand, several studies have presented frameworks for developing native apps. Usually, such frameworks have an editor that enables users to design interfaces, control the behavior of mobile apps, and generate native apps (codes or executables). Lartorre et al. presented a comparative survey of webapp authoring tools for e-learning [12]; however, these tools are specialized for specific purposes. For general purposes, authoring tools such as mbizmaker and app cooker are useful [14], [15]. However, apps developed using these tools have restricted forms and functions predefined by the tools..

3 ARCHITECTURE OF MOBILE E-LEARNING APPLICATIONS

The architecture of a mobile e-learning app may differ greatly depending on the use of mobile web and server content: that is, portions of the native application components may vary between applications. In other words, an e-learning app consists of learning material, learning logics, and a UI, which are referred to as the data, logic, and UI layers in Figure 1, respectively. The data/content layer includes learning objects such as images, files, and video clips. The logic layer contains user logics and learning logics such as navigation events and user activities as well as their overall structure. The UI includes visual and control objects as well as their event-handling mechanisms.

We can construct a so-called stand-alone native app—an app that contains all of these components on a local device (Figure 1(c)). At the other extreme, we can create an app by using a web site and placing a link on a mobile device. Generally, developing and reusing online content is known to be easy with a web-centered architecture. Alternatively, a native component–oriented architecture may work more stably once installed because it contains most of the resources locally. Figure 1 shows the types of mobile application architectures along with their relationships to web content.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

The corresponding functions of the native app modules are as follows:

1) Mobile web page with a light app

As shown in Figure 1(a), Mobile apps of this type usually have only a few functions: showing an opening view, connecting to the web page, and occasionally presenting important notices or management information before commencing the learning sessions. Once connected, the app only performs certain tasks such as calculating statistics and monitoring.

2) Native app with open API (Figure 1(b))

Mobile apps of this type not only have UI functions but also their own interface and learning logics. An open API is a predefined set of APIs that are published by the server and which can be invoked from any client system. Often, the learning material remains on the server with which the client interfaces via open APIs. Some of the learning logics may also reside on the server. In this type of architecture, the app becomes quite complicated because it contains most of the UI and logic layer functionalities locally and must communicate with the server intensively.

3) Stand-alone native app (Figure 1(c))

Native mobile apps contain not only the UI and logic layers but also the data layer at local devices. The local database may hold learning objects such as resources and files as well as the navigation logics and content. Often, this type of native app connects to the server to retrieve a larger amount of content, beyond that which can be accommodated in the mobile device. Therefore, the issue of caching may increase the complexity of the apps.

For all three types, the UI is the essential component of mobile apps. Moreover, meeting the higher level of user expectations is one of the greatest challenges of current mobile e-learning development.

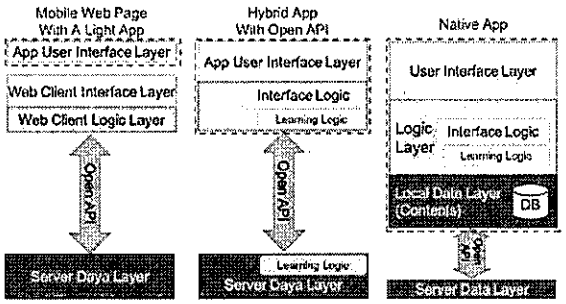


Figure 1. Types of mobile app architecture. .The dashed boxes show the native part of the clients.

4 DESIGN OF PLATFORM-INDEPENDENT UI AUTHORING FRAMEWORK

Although well-established development environments for mobile applications exist (e.g., XCode for iPhones [16], Eclipse for Android [17]), many efforts have been made to provide platform-independent development tools. These approaches are categorized into two groups: those that use standardized web technology, and those that use common UI representations created by editors. Popular examples of the former approach are PhoneGap [18] and Appcellerator [19]. In this paper, we focus on the latter approach in which a framework usually consists of an editor, an IR, and a mobile runtime engine. After authors design the UI and the necessary behaviors, the design result is stored in the form of an IR. This representation is then parsed by the platform-dependent mobile runtime engine to generate the necessary resources and program codes for the target platform. Therefore, the framework consists of the front-end, IR, and back-end runtime engine (Figure 2). The back-end runtime engine is the key to handling platform dependencies.

A. Platform-Independent UI Authoring

In this section, we consider the design of the front-end, an editor subsystem that provides user design activities. The components of the

ISBN: 978-1-941968-02-4 ©2014 SDIWC

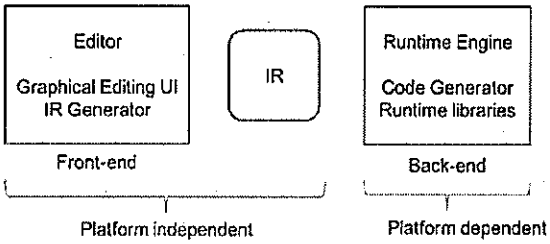


Figure 2. Conceptual diagram of mobile multi-platform authoring framework.

UI are objects, views as containers of objects, and events as well as the behaviors they trigger. In fact, because the UI for e-learning applications does not greatly differ from those of general mobile software systems, this section discusses mobile UI design in general. Platform-independent UI design requires a type of editor that supports neutral feature editing and outputs the design results into IR data. We consider some issues for these components and the overall UI authoring framework.

Although static rendering data for an UI object are well defined, the dynamic behaviors of objects (i.e., animation and event handling) are difficult to design and render on different platforms. However, a qualitative UI requires a certain amount of dynamic behaviors for providing user interactions. The following issues should be considered in constructing an authoring editor for designing mobile UIs.

1) Object properties

UI objects might be images, drawables, text, and multimedia objects. To obtain platform-independent rendering data, the author determines the file location, size, and other properties. These properties are well standardized in web technologies. A view may be regarded as an object in which other objects are embedded with layouts and structures.

2) Animation

Animations enable enhanced learning because they add emotional interactions in addition to increasing the attractiveness of an interface. However, authoring animations requires a large amount of intensive work and proper support of the running platform. This makes authoring animations for mobile apps more challenging.

Active efforts to use animations on mobile platforms are still ongoing. Each platform provides its own animation framework (e.g., CoreAnimation in iOS [20], drawable animations in Android [21]). Several frameworks are available for platform-independent animation authoring; these include, among others, openGL[22], Flash, and animated CSS [23]. Although its efficiency and richness are restricted compared with those of a proprietary framework, openGL is one of the mature open approaches available in various mobile platforms. Therefore, authoring UI designs for animations requires that the underlying animation framework of the target runtime environment be determined in advance.

One possible approach is to use a sprite mechanism, which implements animations by a sequence of frames, to compose several images and drawables. The simplicity and intuitiveness of the sprite approach allows both platform independence and expressive freedom to designers at the cost of user design efficiency and rendering overheads. For sprite animations, the editor interface should provide a way to edit frames and image fractions as well as their sequences.

3) Dynamic event behaviors

Events and the actions they trigger enable dynamic behaviors along with animations. An authoring tool should provide ways to define events for each object and event-triggered reaction by linking events and their handlers. There are two types of events: intra-view and inter-view navigation events. The former includes rendering or user action events that cause reactions within a view. Navigation events are an important means of structuring learning sequences and logics.

Although the mechanisms of event transmission and triggering vary widely between platforms, general event types and their handling mechanisms are well defined by web technology standards.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

B. *Intermediate Representation*

The IR format for outputting the design results of the front-end is crucial for the framework design and performance of the back-end. The static properties of text and images are represented by CSS (Cascade style script) []. A unique identifier should be assigned to each object, event, and view. Identifiers are an important means of representing various objects in UIs; they connect design objects to the program code and perform mapping between separate modules of the framework and mobile app components. The front-end editor creates a set of identifiers and their unique values. In contrast, animations are represented differently depending on the underlying solution. Representations for Open GL and/or flash style animations are available. When the framework determines a specific solution, the corresponding representation is selected.

The dynamic behaviors of a client mobile app are defined by the event mechanism. For this purpose, various representation languages to describe UI behavior have been developed. Most of these are based on XML format, well-defined, and extensible. Adopting a standard representation burdens the back-end runtime engine to support the whole representation language.

Once the IR format is defined, the front-end editor generates the IR data in a straightforward manner. The data generated as output should be parsed in at the beginning of the back-end subsystem.

C. *Design of a Runtime Rendering Engine*

The approach described thus far leaves platform-dependent issues to the back-end engine to liberate the front-end. In general, for the frameworks considered here, most platform-specific functions at the back-end are handled either by the code generated by the framework engine or by linking predefined runtime libraries.

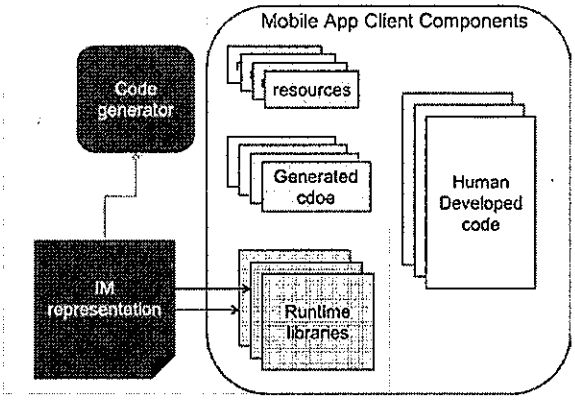


Figure 3 Mobile runtime engine architecture and the mobile app client structure.

Figure 3 shows the typical architecture of the runtime rendering engine subsystem. The code generator module generates resources and data representations in a platform-dependent format. The format may differ between platform operating systems or development languages. In addition, some program codes are generated for the views designed at the front-end. Because a view is a unit of visual structure as well as a structure for navigation and control, the way to handle views depends on the language and platform.

The overall program may be generated automatically if the structure is simple and predictable. One example of such a program is app books, which includes the expected sequential structure of pages, where navigation events have simple predefined forms and a page includes a number of objects and event handlers.

As shown in Figure 5, a set of manually developed codes should be integrated with the generated code by a set of identifiers. A set of to-be-defined identifiers is also generated by the code generator. The manually developed codes should refer to this information to properly map the objects designed at the front-end.

However, in cases where most of the rendering and running mechanisms might be determined beforehand for each platform,

ISBN: 978-1-941968-02-4 ©2014 SDIWC

predefined libraries are superior to code generation. In particular, the implementation of animations should be supported by runtime libraries that are a part of the framework.

Several problems must be solved to support multi-platform authoring. First, the screen size and rendering mechanism must be adapted to each rendering object or view. Generally, using bitmap representations and point unit sizes, the IR achieves platform-neutral representations. The second issue involves event handling, because the framework may adopt a standard representation language or define its own IR format, which the back-end runtime system interprets to generate the corresponding native event processing codes. The last problem concerns animation, which was discussed in the previous section.

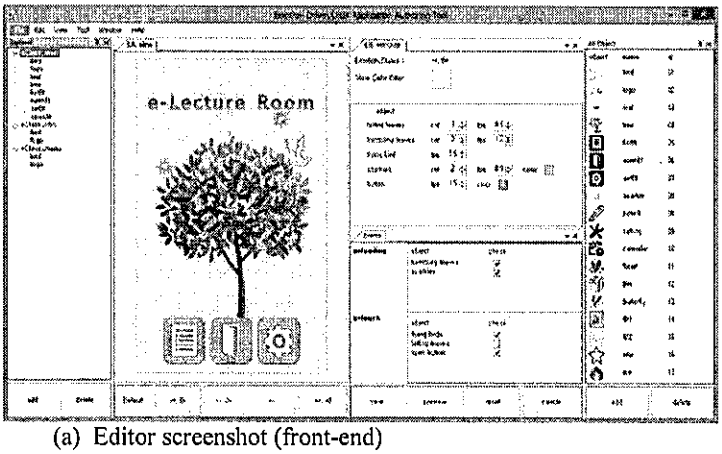
How these platform-dependent issues are solved for each platform, and how the architecture of the generated codes, runtime libraries, manually developed codes, and the integration between them are designed will depend on the implementation.

4 PROTOTYPE IMPLEMENTATION

DAT4UX is an emotional multi-platform UI/UX authoring tool currently being developed by our research group. The motivation for developing the system is to provide a tool for designers to author mobile apps without the help of developers. Because the cost of UI design and development for mobile apps is excessive compared with that for web or desktop software, we decided to develop a design tool with which designers could develop qualitative interfaces on their own. We developed a framework following the front- and back-end architecture introduced in the previous section. Moreover, we focused on the ease of integration between automatically generated modules and manually developed codes. Another design goal of the framework was to provide unrestricted design functionalities for dynamic visual effects that enable emotional user experiences. To achieve these goals, we developed a front-end desktop editor subsystem and back-end runtime rendering engines for and Android platforms.

The ability to adapt the screen size and rendering mechanism to each rendering object or view should be supported to enable multi-platform authoring. By using bitmap representation and point unit size, the IR provides neutral representations. In addition, the target screen size and corresponding preview can be set with the editor.

The front-end editor is implemented in Visual Studio 2012, using C# language. We selected C# to guarantee performance of image processing, because implementing frame animations requires significant computing speed. We adopt a sprite-style animation



(a) Editor screenshot (front-end)

```
- <SceneList>
  - <Scene Id="100" Name="Intro" Width="720" Height="1280
    - <Frame FrameNo="0" Delay="1">
        <Layer ImageId="2" X="200" Y="300" FilTer="Normal"
        <Layer ImageId="3" X="120" Y="240" FilTer="Normal"
        <Layer ImageId="4" X="420" Y="690" FilTer="Normal"
      </Frame>
    - <Frame FrameNo="1" Delay="1">
        <Layer ImageId="2" X="220" Y="320" FilTer="Normal"
        <Layer ImageId="3" X="120" Y="250" FilTer="Normal"
        <Layer ImageId="4" X="430" Y="700" FilTer="Normal"
      </Frame>
    </Scene>
  - <Scene Id="200" Name="MainMenu" Width="720" Height="
    - <Frame FrameNo="0" Delay="1">
        <Layer ImageId="22" X="10" Y="10" FilTer="Normal" Bl
        <Layer ImageId="30" X="650" Y="10" FilTer="Normal"
      </Frame>
    - <Frame FrameNo="1" Delay="1">
        <Layer ImageId="22" X="10" Y="10" FilTer="Normal" Bl
        <Layer ImageId="30" X="650" Y="10" FilTer="Normal"
      </Frame>
    </Scene>
```

(b) IR codes (part)

Figure 4. Running Examples of Prototype Implementation System DAT4UX

ISBN: 978-1-941968-02-4 ©2014 SDIWC

mechanism to allow abundant and unrestricted animation effects, as shown in Figure 4(a). Authors enjoy the Power-Point–style editing tool, which supports various animation effects. To support richer and more active dynamic effects, we are considering the open GL animation platform.

We defined our own XML format for IR to implement the prototype. The IR includes information for representing the design results of the editor in a platform-independent manner, as shown in Figure 4(b).

The back-end runtime engine of DAT4UX consists of a code generator and the runtime engine. The code generator creates resource files from the IR information, and the runtime engine consists of platform-dependent libraries that render the views and visual effects. User-developed codes can use the libraries to show the view. The mobile app project would include resources and libraries.

5 CONCLUSION

As mobile e-learning has propagated, it becomes more important to develop mobile user interface in e-learning system development practice. However, there are difficulties from multi-platform and platform dependence issues in mobile computing environment. We surveyed current approaches for these problems including web- or native-app based frameworks.

In this paper, we proposed a design of multiplatform native app development frameworks consisting of front- and back-end subsystems. The front-end subsystem is an editor which enables designers to author the user interface without help of developers. Also, the editor would emit intermediate representation of the design result, which is used as rendering information for the back-end. The back-end system provides runtime environment for the specific platform, which consists of a code generator and runtime libraries.

The proposed design is implemented as a prototype system called DAT4UX. This system is a multi-platform, codeless, qualitative interface authoring tool, with which designers can work with animation or dynamic effects of user interface without developer's help. We presented implementations for Android platform. We have a plan to extend the system with enhanced functionalities and usability while preserving the advantage of general architecture.

6 ACKNOWLEDGEMENTS

This research was supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the Specialized Cooperation between industry and academic support program
(NIPA-2014-H0806-14-1004) supervised by the NIPA(National IT Industry Promotion Agency)

7 REFERENCES

[1] A. Charland, and L. Brian, "Mobile application development: web vs. native." *Communications of the ACM* 54.5 (2011): 49-53.

[2] I. Dalmasso, S.K. Datta, C. Bonnet and N. Nikaein, "Survey, comparison and evaluation of cross platform mobile application development tools." *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*. IEEE, 2013.

[3] M.E. Joorabchi, A. Mesbah, and P. Kruchten, "Real challenges in mobile app development." Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on. IEEE, 2013.

[4] F. Shull, "Designing a World at Your Fingertips: A Look at Mobile User Interfaces." *Software, IEEE* 29.4 (2012): 4-7. 071

[5] D. Guralnick, "Designing effective e-learning user interfaces." IADIS International Conference on the WWW and the Internet. October 5-8, 2007, Vila Real, Portugal. 2007.

[6] S. Levy and J. Yupangco, "A picture is worth 1000 words: Visual design in e-learning." Learning Solutions: Design Techniques (2008): 1-8.

[7] J. Potts, N. Moore and S. Sukittanon, "Developing mobile learning applications for electrical engineering courses." Southeastcon, 2011 Proceedings of IEEE. IEEE, 2011.

[8] Y. Park, "A pedagogical framework for mobile learning: Categorizing educational applications of mobile technologies into four types." *The International Review of Research in Open and Distance Learning* 12.2 (2011): 78-102.

[9] M. Sharples, J. Taylor and G. Vavoula, (2007) "A Theory of Learning for the Mobile Age." In R. Andrews and C. Haythornthwaite (eds.)The Sage Handbook of E-learning Research. London: Sage, pp. 221-47.

[10] M. Virvou, and E. Alepis, "Mobile educational features in authoring tools for personalised tutoring." *Computers & Education* 44.1 (2005): 53-68.

[11] A. Gordillo, E. Barra, D. Gallego, J. Quemada, "An online e-Learning authoring tool to create interactive multi-device learning objects using e-Infrastructure resources." *Frontiers in Education Conference, 2013 IEEE*. IEEE, 2013.

[12] M. Latorre, A. Robles-Gomez, L. Rodriguez, P. Orduna, E. San Cristobal, A.C. Caminero, L. Tobarra, I. Lequerica, S. Ros, R. Hernandez, M. Castro, D. Lopez-de-Ipina, J. Garcia-Zubia, "A review of webapp authoring tools for e-learning." *Global Engineering Education Conference (EDUCON), 2014 IEEE*. IEEE, 2014.

[13] A. Zibula and T.A. Majchrzak, "Cross-Platform Development Using HTML5, jQuery Mobile, and PhoneGap: Realizing a Smart Meter Application." *Web Information Systems and Technologies*. Springer Berlin Heidelberg, 2013. 16-33.

[14] App Cooker, http://www.appcooker.com/

[15] m-BizMaker - world best mobile platform, http://www.mbizmaker.com/ups/mbizmaker/index.html

[16] Xcode, https://developer.apple.com/xcode

[17] Android Development Tools (ADT), http://developer.android.com/tools/sdk/eclipse-adt.html

[18] PhoneGap Home, http://phonegap.com/

[19] Appcelerator, http://www.appcelerator.com/

[20] Core Animation Programming Guide, https://developer.apple.com/library/mac/documentation/Cocoa/Conceptual/CoreAnimation_guide/Introduction/Introduction.html

[21] Drawable animation for Android, http://developer.android.com/guide/topics/graphics/drawable-animation.html

[22] OpenGL, https://www.opengl.org/

[23] CSS3 animations, http://www.w3schools.com/css/css3_animations.asp

Shifting Virtual Reality Education to the Next Level – Experiencing Remote Laboratories through Mixed Reality

Max Hoffmann, Tobias Meisen and Sabina Jeschke
Institute for Information Management in Mechanical Engineering, RWTH Aachen University
Dennewartstrasse 27, 52068 Aachen, Germany
max.hoffmann@ima.rwth-aachen.de

ABSTRACT

Technical universities are more and more focusing on engineering education as a primary discipline. All along with the integration of various innovative fields of application into the curriculum of prospective engineers the need for appropriate educational features into the studies also increases. Unlike exclusively theoretical studies as physics, mathematics or information sciences the education of engineers extensively relies on the integration of practical use-cases into the education process. However, not every university is able to provide technical demonstrators or laboratories for all of the various applications in the field of engineering. Thus, it is the aim of the current paper to propose a method that enables visiting a high variety of engineering laboratories based on Virtual Reality. A Virtual Reality simulator is used to create and emulate remote laboratories that can be located at arbitrary places far away from their Virtual Reality representation. This way, by melting real world demonstrators with virtual environments, we enable a physically and technically accurate simulation of various engineering applications. The proof of concept is performed by the implementation and testing of a laboratory experiment that consists of two six-axis robots performing collaborative tasks.

KEYWORDS

Virtual Reality, Mixed Reality, Augmented Virtuality, Virtual Learning Environments, Remote Laboratories, Engineering Education.

1 INTRODUCTION

In a world that is increasingly based on scientific innovations and technological progress the education of engineering students constantly gains in importance. Furthermore, the field of studies and the variety of possible specializations in engineering classes also increase significantly. In order to qualify graduated engineers to enter the working world successfully, the imparting of practical knowledge during studies plays a decisive role in education.

While theoretical knowledge is still transferred using written texts and the spoken word – normally through lecture notes and traditional readings – the experience of practical use-cases in engineering education commonly relies on the visit of laboratory experiments during the studies. However, since computer vision and digitalization techniques have grown to an extensive level, the integration of new media into the curriculum replacing the visit of real laboratories gains in importance [1]. Especially in terms of engineering applications, the use of high quality and realistic visualization techniques as a supplement to the attendance within practical laboratory experiments is of major importance for successfully impart basic concepts of engineering applications to students. Not at least due of the progress in computer science and graphical visualization techniques, the capabilities of visualizing objects of interest embedded into an artificially designed context have grown to an exhaustive amount. In this context, for example physical effects or technical subtleties of engineering applications can be presented in higher detail or in an amplified way in order to emphasize aspects that are not easily observable in reality. These novel potentials can be utilized to

ISBN: 978-1-941968-02-4

explain theoretical knowledge more concrete and tangible and help engineering students to understand the concepts of complex technical applications on different levels and from a practical point of view.

Another major trend that is emerging within the field of education and learning relies on the way of distributing information and knowledge through internet-based media among the students. The high significance of online platforms and social media during the everyday life of a student can be exploited in order to increase communication channels by the establishment of E-Learning-Platforms [2]. The technical possibilities of sharing and representing educational contents, spreading knowledge over the world-wide web and enabling the remote participation of students in engineering classes, open up new opportunities of teaching and learning within universities. In combination with modern visualization techniques these computer-based teaching concepts can for example be implemented through Virtual Reality or Mixed Reality applications [3].

Despite all these technological possibilities, universities are facing more and more obstacles to deal with the high variety of engineering courses and their different technical applications and needs. Thus, each university can only provide a limited amount of experimental or laboratory classes for the different fields of engineering during their curricula. However, the demand for enlarging the variety of engineering education contents is also growing constantly. This leads to a conflict of interests as universities are not supposed to advantage any particular engineering class in comparison to the others. Due to the limited laboratory and teaching capacities, the ability to satisfy the needs for experimental education of engineering students' cannot be fulfilled adequately. Furthermore, most of the offered experimental simulations or virtual representations of laboratory environments are lacking the required quality standards in terms of graphical accuracy and interaction capabilities.

The goal of the current paper is to find ways for dealing with these obstacles and proposing possibilities to enable an extensive practical education of engineering classes by answering the following questions that are correlated to the described issues:

1. How can universities address the high variety of engineering related disciplines by ensuring the availability of suitable use-cases, experiments and laboratory exercises for students?
2. How is it possible to integrate concepts of Experiential Learning into engineering education against the background of limited laboratory resources?
3. What is the benefit of integrating Virtual Reality and Mixed Reality applications into engineering education from the technical/physical point-of-view?
4. Is it possible to address educational needs of engineering students in universities by enabling an active manipulation of the virtual laboratory environments?

The present publication intends to answer these questions by introducing a novel method for realizing use-cases for Virtual Reality exercises and laboratory experiments in terms of Remote Laboratories. In this context, the term Remote Laboratory introduces the idea of enabling the visit of distant places for conducting laboratory experiments based on its virtual representation. Hence, it is our aim to enable students to attend specific experimental environments suitable to their field of studies, even if these exercises are not provided at their particular university.

In the next section, the state of the art in teaching with the aid of new media and novel visualization applications is presented. Furthermore, educational advantages of Virtual Reality and Mixed Reality applications are carried out. In section 3 the technical implementation of our next-generation Virtual Reality simulator is described together with former virtual scenarios that have been implemented using the simulator. Section 4 presents a scenario for enhancing

ISBN: 978-1-941968-02-4 ©2014 SDIWC

students' learning behavior by the creation of a Remote Laboratory in connection with Mixed Reality approaches. In section 5 first attempts in evaluating the advantages of such virtual representation of a laboratory experiment are carried out, assessing the learning capabilities of a group of students visiting such Remote Laboratory. Section 6 concludes the outcome of the current paper and specifies the next steps of enhancing the user's experience and creating larger Remote Laboratory environments.

2 STATE OF THE ART

New media have gained high significance in university studies and publications in the past decade. These new media – which are mostly based on computer visualization techniques – are continuously replacing traditional books and lecture notes. White boards and projectors are replaced by presentation software, which represents the new standard for visualizing text and pictures [4], with PowerPoint as market leader [5]. This switch from traditional lecture speech to graphical representation has been performed, because this form of presentation enables focusing on the main points of the educational content using illustrative representations and pictorial summaries [6]. Despite the positive, but also critical discussion about an overwhelming usage of PowerPoint [7–9] as primary teaching tool, the usage of presentation software in the classroom has grown constantly [10, 11]. In connection with the entry of technological novelties into the classroom it is time to take the next steps from merely presenting pictures using presentation software to the usage of advanced graphical interfaces opening up interaction capabilities for the students involved in the engineering courses.
Presentation software like PowerPoint may be a far reaching advancement for most courses of studies. However, the usage of these meanwhile basic IT tools is also limited to a certain kind of knowledge transfer. Especially for practically oriented study paths like engineering classes active interaction capabilities within courses and exercises are inevitable. In these highly technical studies there is an urgent need for interactive laboratory experiments in order to impart practical and skill-based knowledge tangible to students. Against this background, David A. Kolb's traditional, well-established cycle on Experiential Learning is more up to date than ever [12]. The – almost classical – Learning Cycle is depicted in Figure 1.

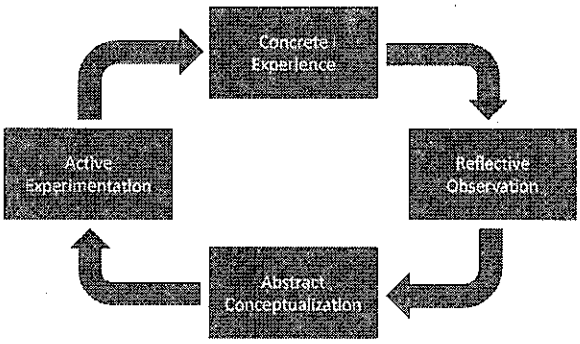


Figure 1. The Experiential Learning Cycle according to David A. Kolb (1984)

In the picture, we see a never ending process of active experimentation, concrete experience, reflective observation and abstract conceptualization. Starting at active experimentation, there is a need for concrete experience in order to understand abstract concepts. The reflective observation following to the experience helps carrying out an abstract conceptualization that is based on a deeper understanding of the experienced content. Especially the practical part of the learning process in terms of the attendance to experimental courses cannot be replaced by any kind of theoretical knowledge transfer. Other learning related theories, which address the same matters as the experiential learning approach, are action learning, adventure learning, free choice learning, cooperative learning and service learning approaches [13]. In all of these theories, active interaction of the learning person plays an integral part in the learning process.
Due to the high complexity and advanced technological level of the relevant applications in engineering classes, sole static visualizations

ISBN: 978-1-941968-02-4 ©2014 SDIWC

are not capable to serve as a medium for active experimentation or even concrete experience. In order to address these parts of the learning cycle, novel visualization concepts have to be applied. A key enabler that combines advanced visualization techniques with the experience of a certain scenario is Virtual Reality as shown in the relevant literature [14] and during prior studies in this field of application [15]. In the first step, Virtual Reality cannot serve direct interaction possibilities, but through the use of immersive effects, a Virtual Reality scenario is able to arouse an experienced reality within the perception of the user. This effect can be greatly characterized by the definition of immersion according to Murray [16]: "Immersion is a metaphorical term derived from the physical experience of being submerged in water. We seek the same feeling from a psychologically immersive experience that we do from a plunge in the ocean or swimming pool: the sensation of being surrounded by a completely other reality, as different as water is from air that takes over all of our attention, our whole perceptual apparatus."

In order to address the part of concrete experience during practical education accurately by virtual reality applications, the utilized tools have to fulfill a number of conditions to serve as a suitable complement for laboratory classes. In the following, we will therefore look at existing Virtual Reality applications and discuss if these are capable of bridging the gap of concrete experience to the students.

In terms of existing Virtual Reality technologies, there are already many technical solutions that are primarily focused on the creation of high-quality and complex three-dimensional environments, which are accurate to real-world scenarios in every detail. One example are flight simulators that are capable of tracking the locomotion of a flying vehicle in a virtual scenario [17]. However, these systems are usually not taking into account the position or the head movements of the user. Another Virtual Reality simulator is the well-known Omnimax Theatre, which provides a large angle of view [18], but does not allow any tracking capabilities whatsoever. First attempts to interact with Virtual Reality in a natural way were introduced by head-tracking monitors as conducted by Codella [19] and Deering [20]. These specially designed monitors provide an overall tracking system, but are characterized by a rather limited angle of view [17]. The first mentionable approach to create a Virtual Reality environment with full tracking capabilities of movements and of the head position of the user was introduced by McDowall [21] with the Boom Mounted Display. Despite advanced tracking capabilities these early attempts were characterized by poor resolutions and thus were not capable of a detailed graphical representation of a virtual environment [22].

Thus, in order to enable true user experience in simulated scenarios, Mixed Reality approaches have to be embedded into the Virtual Reality, were reality and virtuality are merged into each other [23]. One far reaching innovation in terms of enabling Virtual Reality and Mixed Reality applications was introduced with the CAVE in 1992 by Cruz-Neira et al. [24]. Hereby, the recursive acronym CAVE stands for Cave Automatic Virtual Environment. By making use of complex visualization techniques combined with various projectors and six projection walls arranged in form of a cube, the developers of the CAVE have redefined the standards in visualizing Virtual Reality scenarios by enabling a new level of immersion.

The CAVE reaches further towards true Virtual Reality which – according to Rheingold [25] – is described as an experience, in which a person is "surrounded by a three-dimensional computer-generated representation, and is able to move around in the virtual world and see it from different angles, to reach into it, grab it and reshape it." These active manipulation activities that can be performed by the user open up various new applications in education by rebuilding industrial use-cases. Thus, by enabling extended interaction capabilities with the scenarios in terms of providing relatively free manipulation of the virtual environment,

ISBN: 978-1-941968-02-4 ©2014 SDIWC

the immersive effects of the scenarios are enhanced. This effects could lead to the desired impact on the learning behavior of students that consists of the ability to derive abstract conceptualizations on the basis of concrete or practical experience and active experimentation.
However, even the CAVE has got restricted interaction capabilities as the user can only interact in the currently demonstrated perspective of the scenario. Furthermore, natural movement is limited, as locomotion through different scenes of the scenario is usually performed by flying to the next spot. Yet, natural movement as walking, running or jumping through the Virtual Reality is decisive for a highly immersive experience in the virtual environment.
In order to fill this gap of limited interaction and accordingly deeper immersion into the scenario, additional devices and tracking systems for allowing such interaction have to be included into the scenario without losing the high quality of graphical representation of the virtual environment. One promising approach relies in the establishment of the Virtual Theatre that brings together a full-size stereo-vision view and various interaction devices and manipulation capabilities for the user.

3 THE VIRTUAL THEATRE – ENABLER FOR EXTENDED IMMERSION

The Virtual Theatre represents a next-level Virtual Reality simulator that allows free locomotion in a virtual environment and active manipulation as well as the control of objects in the visualized scenario. The Virtual Theatre was carried out by a Swedish company [26] and was already described in detail in our previous publications [27] according to the scenarios that have been carried out during our latest research [28, 29]. The centerpiece of the Virtual Theatre is the omnidirectional treadmill, a moving floor that accelerates its centric arranged rollers according to the position of the user (Figure 2).

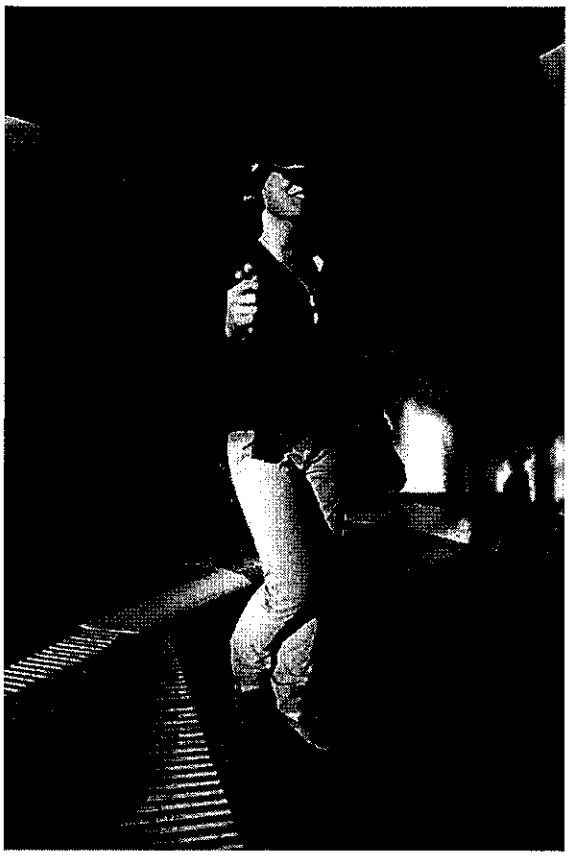


Figure 2. A user experiencing virtual environments in the Virtual Theatre

The user himself is tracked by an infrared tracking system; hence his head and hand movements are constantly observed and taken into account for a adaptation and manipulation of the virtual environment. The user is wears a Head Mounted Display (HMD) that is equipped with two screens – one for each eye – and enables highly immersive three-dimensional stereo vision for the exploration of the virtual space. The unique characteristic of a HMD is that this kind of devices are capable of measuring the user's head orientation through a perpendicular, which makes it possible to adjust the Virtual Reality according to the head's actual position and orientation. This enriches the immersive experience of the user as he is able to look around and explore the Virtual Reality in a similar way as he does in the real world. An embedded sound system into the

ISBN: 978-1-941968-02-4 ©2014 SDIWC

HMD completes the plunge into virtuality. For further information about the technical concept of the Virtual Theatre the reader is encouraged to refer to [28], where the hardware and technical setup is explained in detail.

Former scenarios that have been carried out using the Virtual Theatre were well received by the students of engineering classes [29]. One example for the creation of a huge sized virtual environment is our Mars project, where an extensive simulation of a plateau on the surface of the red planet was recreated to enable upcoming astronautics and aerospace students to perform a virtual visit and exploration of the Mars [15]. Another application of the Virtual Theatre was carried out in terms of a study in order to assess the learning behavior and learning efficiency of students while being surrounded in a virtual environment [29]. During the survey, the students were located in a virtual labyrinth, in which they needed to find objects and recognize their location and shape at a later point. Afterwards, the results of their learning efficiency were evaluated and compared to the efficiency using traditional computer screens for performing similar tasks.

However, former studies did not take into account deeper interaction capabilities as an active movement of objects or the remote control of devices for industrial use-cases. However, as mentioned earlier, exactly these interaction capabilities are strongly needed in order to create realistic virtual representation of experimentations, thus to enable Remote Laboratories. Hence, in the next step we present the inclusion extensive interaction capabilities into the use-case and we attempt to enable true immersion based on Mixed Reality concepts.

4 ENABLING REMOTE LABORATORIES THROUGH MIXED REALITY

As part of a network of administrative computers the Virtual Theatre – including its integrated parts and the tracking system – can be expanded by additional hardware in order to enable a more natural user interaction within the virtual scenario.

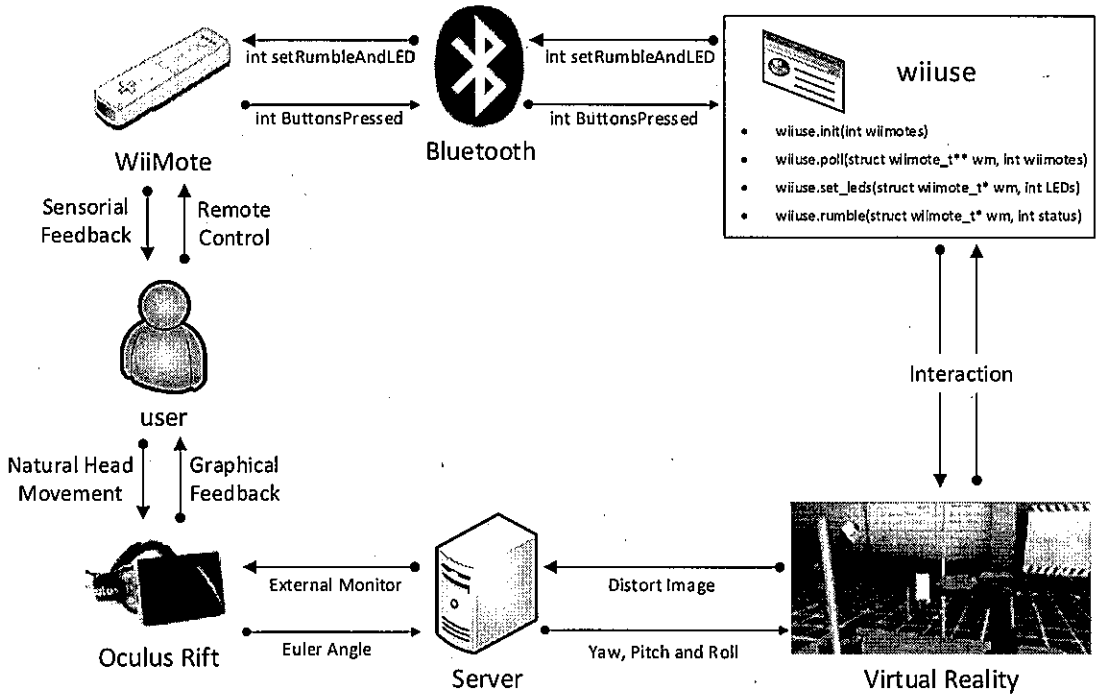


Figure 3. Communication Infrastructure of the Virtual Theatre with extended interaction capabilities

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Taking into account natural user behavior as well as the experience of students using well-established computer game devices, our team decided to carry out a remote control for virtual scenarios based on hardware of the ©Nintendo Wii™ Controller. The new conceptual design of the communication infrastructure for the Virtual Theatre and its surrounding hardware equipment is depicted in Figure 3.
In the middle part of the picture's bottom, the central server is visualized. The server deals with the signals of the Head Mounted Display, which is located on the left side, and processes its information for the user's movement, head position and orientation according to the virtual environment that is depicted on the right side at the bottom. The Wii™ remote controller is connected via Bluetooth and sends specific commands to the central server. The server processes these commands to manipulate the virtual environment and visualize the modified scenario in real-time.
A suitable application for including the described interaction device into a virtual scenario is based on an extended laboratory experiment that was virtualized by our team for education purposes. The setup consists of two six-axis robots that are placed on a table in order to perform collaborative tasks. The virtualized model is depicted in Figure 4.

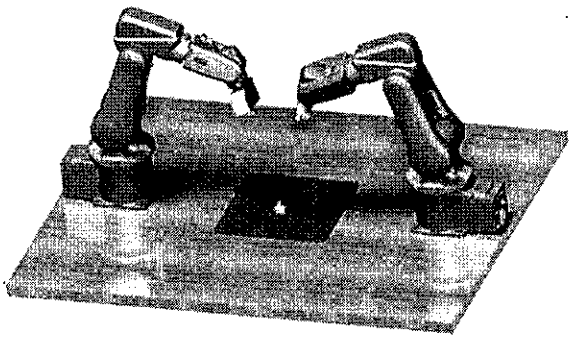


Figure 4. Visualization of two six-axis ABB™ robots performing collaborative tasks

In reality, the robots are located in the same distance as illustrated in the figure, enabling the ability to perform collaborative, interdepending tasks. As described in [28], the first attempts in carrying out a Remote Laboratory based on these robots consisted in the virtualization of the actual robot movements as well as a real-time alignment of the movements that were performed by the real world robots and the movements of our simulation. As shown in our previous publication, the full setup can be appropriately simulated in real time, i.e. the user inside of the virtual reality simulator can pursue the robot motion without any perceptible time lag. This real-time synchronization between the real world laboratory and its virtual representation enables active remote control of the robot setup as well as various use-cases:

- The actual position of the robots can be tracked and remotely manipulated from arbitrary locations.
- The control of the robot arms can be extended by additional security layers in order to assure save motion of the robots and to avoid collisions.
- The experimenter can easily work with dangerous materials or substances (e.g. chemicals) and is able to operate the robots if these are located at dangerous or non-accessible places.

In order to generate an added value to the remote control of a laboratory environment our research did not only concentrate on the development of remote control devices. We also focused on providing additional features that are enabled through Virtual Reality. One major progress of these efforts relies on the integration of Mixed Reality elements into the laboratory experimental context. In this connection, the term Mixed Reality is characterized as the merging of real and virtual worlds to produce new environments and visualizations, where physical and digital objects co-exist and interact in real-time [30]. Our use-cases including Mixed Reality approaches is able to address several aspects of the application:

ISBN: 978-1-941968-02-4 ©2014 SDIWC

- Systematic simulation of experiments with actual machines and components in real time. The exact simulation leads to a co-existence of real world objects in the laboratory and in Virtual Reality with interdepending system states.
- Feedback and manipulation capabilities of the user, which leads to an interaction with both, the objects in the virtual environment and the real components.
- Embedding of real world features into virtual environments by placing cameras into the laboratory environment. This enables the projection of detailed views of experimental insights that are captured by the camera onto a wall in the virtual environment. In terms of the Mixed Reality concept, this effect is also referred to as *Augmented Virtuality*.

Especially the last of the mentioned points of integrating Mixed Reality concept into our Virtual Reality scenarios bears a high potential to enhance the grade of immersion signifycantly. In terms of the user's perception the embedding of Augmented Virtuality – i.e. real pictures into the virtual scenario – is connected to effects of fuzziness between reality and virtuality, which leads to highly immersive impressions for the user. One possible application of integrating a camera in the real world scenario can for example consist in a placement of the camera in a bird's-eye perspective on top of the experimental setup to see the whole scene from a broader point-of-view.

Another possible scenario is to attach the camera onto a robot that is actually moving and thus to observe the scene from the robot point of view. An example for this extended perspective based on this Augmented Virtuality visualization technique is depicted in Figure 5.

On the bottom of the picture the simulation of the two six-axis robots is shown. The screen that is located above the table shows a video image that is taken from the perspective of the left robot in reality. This image is embedded into the Virtual Reality scenario as a sort video projection and hence represents an Augmented Virtuality element in the simulation.

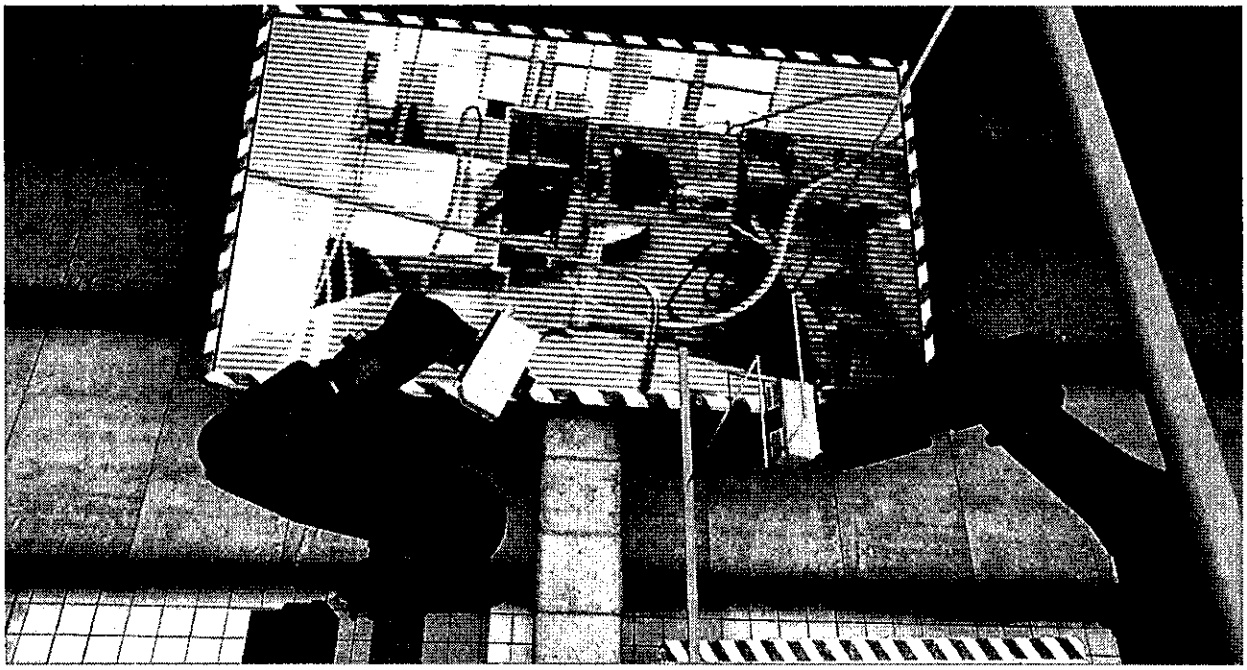


Figure 5. The experimental setup in reality visualized in a Virtual Reality scenario

ISBN: 978-1-941968-02-4 ©2014 SDIWC

This enrichment of the virtual scenario is connected to several improvements concerning the technical and the educational application of the scenario:

- The point of view perspective enables a detailed view of the simulated scene. Through the robot's perspective the tasks that have to be performed by the robot arms can be conducted with higher precision due to the overview based on more than one perspective.
- The grade of immersion increases as the user of the simulation is able to see reality objects that are melting with the Virtual Reality scenario in real-time.
- For education purposes, multiple cameras can be attached at various locations of the demonstrator, which helps to explain the physical or technical effects of use-cases.

Besides the advantages for the single user that are connected to the embedding of Mixed Reality into the scenario there is also an added value for the remote control of a laboratory experiment from more than one user. Due to the placement of multiple cameras within the surroundings of the experiment, different users can perform collaborative tasks while observing a simulation of the actual system state in real-time, but from different perspectives. This enables a highly precise manipulation of the experimental conditions influenced by different users that can be located at arbitrary places. Especially due to this point, the far reaching benefits of Remote Laboratories as a new class of conducting experiments becomes known.

5 EVALUATION OF THE TECHNICAL IMPLEMENTATION AND IMPACT ON THE STUDENTS LEARNING BEHAVIOR

During the previous sections of this paper the added value of Remote Laboratories has been derived in terms of the overall usability and availability of laboratory experiments as well as the impact on the precise conduction of experiments using multiple information channels. In this section we would like to investigate the impact of the utilization of Remote Laboratories not only in terms of the availability of experimental setups in engineering classes, but also its effects on the learning behavior and motivation of students. In terms of the assessment according to this impact, we get back to initial research questions about enhancing the learning environment of students by remote setups or literally: "Is it possible to address the educational needs of engineering students in universities by enabling an active manipulation of the virtual laboratory environments?"
In order to answer this question, different facets of the educational needs of engineering students are taken into account:

- The impact of virtual experience and the grade of sensed immersion into the student's perception in virtual environments to build up a realistic scenario.
- The physical accuracy of simulations embedded into the virtual environment.
- The added value of enhancing the perspective of the user and the personalized view by emphasizing certain physical effects through amplified physical behavior or close-up views.

The first evaluation phase of these effects was performed in-house and based on the personnel and on the student employees of our institution. The visualization accuracy as well as the immersive effects of the simulation into the perception of the user could be verified during the testing phase. Especially the active motion of the six-axis robots through an easily manageable interface while being submerged in the virtual environment had clearly observable impacts on the understanding of robot motion and the need for automation.
In the next steps of the evaluation phase, the investigation of the scenario will be performed by a questionnaire that is carried out for

ISBN: 978-1-941968-02-4 ©2014 SDIWC

laboratory classes of newcomer students. This evaluation, which will be further concentrating on didactical aspects of the experimentation environment, will take place in the following semester, in which the according students will assess their personal learning success after conducting several experiments with and without the help of the described Mixed Reality-related techniques. During this phase we will examine the effects of the virtual environment on the learning behavior of the engineering students by taken into account the following points:

- The impact of virtual experience and the grade of sensed immersion on the willingness and learning behavior of students in virtual environments.
- Correlations between the learning ability during laboratory experiments, gamification effects and fun in manipulating the laboratory environment in virtuality.
- The effect of hands-on experiments on the learning success of students in comparison to the mere observation of distant experiments that are not accessible in its real environment.

The study will show, if the different aspects like fun in learning, active involvement and free movement in virtual environments as well as the ability to manipulate a virtual representation of a real world demonstrator or – in other words – the reflective observation have a significant impact on the abstract conceptualization of complex engineering applications.

6 CONCLUSION AND OUTLOOK

The need for higher capacities in terms of the practical education of engineering students comprises a major challenge for today's universities. The constantly growing number of students with various educational backgrounds and different experiences as well as the wealth of study opportunities demands for innovative concepts in the organization of a profound engineering education.

In this paper, we have substantiated the idea of conducting real laboratory experiments through a Virtual Reality simulator by enabling Remote Laboratories. These laboratories can serve as an extensive supplement to real experimental setups, because they can be built up at arbitrary places and run simultaneously for multiple users. Various setups for virtual environments can be applied in order to emphasize immersive effects on the user with an expected impact on his learning behavior.

The next steps in connection with the presented scenarios will consist in a quantitative evaluation of the impact of Virtual Reality on the actual learning success of the students by assessing the conceptual knowledge of two different comparison groups, one that visits an actual laboratory experiment without any personal involvement or interaction with actual components, and the other group that visits a Virtual Reality based virtual environment of the laboratory experiment. Furthermore, we will discuss the effect of embedding Mixed Reality components into the Remote Laboratory on the students on the one hand in terms of their qualitative perception of being immersed into the virtual environment and on the other hand in terms of the advantages that are connected to the embedding of camera screens into the virtual scenario for additional perspectives.

On the technical side, the next steps concerning an extension of the Remote Laboratory environment consist in the development of a generic methodology to automate and to control robots of various kinds in virtual environments. In terms of this procedure, aspects of robot security, collision avoidance and inverse kinematics for robot control will be of major importance for an expedient experimentation environment. Next projects will concentrate on the implementation of complex scenarios with multiple robots and interaction devices in order to emphasize the idea of collaborative and concurrent engineering in virtual environments.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

ACKNOWLEDGEMENT

The present work was conducted in terms of the investigation of virtual and remote laboratories in research and teaching and was supported by the project "ELLI – Excellent Teaching and Learning in Engineering Sciences" as part of the Excellence Initiative at the RWTH Aachen University.

REFERENCES

[1] M. Ebner and A. Holzinger, "Successful implementation of user-centered game based learning in higher education: An example from civil engineering", Comp. & Education 49, vol. 3, pp. 873–90, 2007.

[2] M. J. Rosenberg, "E-learning". Strategies for delivering knowledge in the digital age. New York: McGraw-Hill, 2001.

[3] Z. Pan, A. D. Cheok, H. Yang, J. Zhu, and J. Shi, "Virtual reality and mixed reality for virtual learning environments", Computers & Graphics 30, vol. 1, pp. 20–28, 2006.

[4] A. Szabo and N. Hastings, "Using IT in the undergraduate classroom". Should we replace the blackboard with PowerPoint? Computer & Education 35, 2000.

[5] R. J. Craig and J. H. Amernic, "PowerPoint Presentation Technology and the Dynamics of Teaching", Innov. Higher Educ. 31, vol. 3, pp. 147–60, 2006.

[6] R. A. Bartsch and K. M. Cobern, "Effectiveness of PowerPoint presentation in lectures", Computer and Education 41, pp. 77–86, 2003.

[7] T. Creed, "PowerPoint, No! Cyberspace, Yes", The Nat. Teach. & Learn. F. 6, vol. 4, 1997.

[8] D. Cyphert, "The problems of PowerPoint". Visual aid or visual rhetoric?, Business Communication Quarterly 67, pp. 80–83, 2004.

[9] P. Norvig, "PowerPoint: Shot with its own bullets", The Lancet 362, pp. 343–44, 2003.

[10] T. Simons, "Does PowerPoint make you stupid?", Presentations 18, vol. 3, 2005.

[11] A. M. Jones, "The use and abuse of PowerPoint in teaching and learning in the life sciences". A personal view, BEE-j 2, 2003.

[12] D. A. Kolb, "Experiential learning". Experience as the source of learning and development. Financial Times Pren Hall, 2014.

[13] C. M. Itin, "Reasserting the Philosophy of Experiential Education as a Vehicle for Change in the 21st Century", Journal of Experiential Education 22, vol. 2, pp. 91–98, 1999.

[14] D. Johansson and L. J. de Vin, "Towards Convergence in a Virtual Environment: Omnidirectional Movement, Physical Feedback, Social Interaction and Vision", Mechatronic Systems Journal, November 2011.

[15] M. Hoffmann, K. Schuster, D. Schilberg, and S. Jeschke, "Next-Generation Teaching and Learning using the Virtual Theatre", 4th Global Conference on Experiential Learning in Virtual Worlds. Prague, Czech Republic, 2014.

[16] J. H. Murray, "Hamlet on the Holodeck". The Future of Narrative in Cyberspace. Cambridge (Mass.): MIT Press, 1997.

[17] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti, "Surround-Screen Projection-based Virtual Reality. The Design and Implementation of the CAVE", SIGGRAPH '93 Proceedings of the 20th annual conference on Computer graphics and interactive techniques. ACM - New York, pp. 135–42, 1993.

[18] N. Max, "SIGGRAPH '84 Call for Omnimax Films", Computer Graphics 16, vol. 4, pp. 208–14, 1982.

[19] C. Codella, R. Jalili, L. Koved, B. Lewis, D.T. Ling, J.S. Lipscomb, D. Rabenhorst, C.P. Wang, A. Norton, P. Sweeny, and G. Turk, "Interactive simulation in a multi-person virtual world", ACM - Human Fact. in Comp. Syst. CHI 1992 Conf., pp. 329–34, 1992.

[20] M. Deering, "High Resolution Virtual Reality", Com. Graph. 26, vol. 2, pp. 195–201, 1992.

[21] I. E. McDowall, M. Bolas, S. Pieper, S. S. Fisher, and J. Humphries, "Implementation and Integration of a Counterbalanced CRT-based Stereoscopic Display for Interactive Viewpoint Control in Virtual Environment Applications", Proc. SPIE 1256, vol. 16, 1990.

[22] S. R. Ellis, "What are virtual environments?", IEEE Computer Graphics and Applications 14, vol. 1, pp. 17–22, 1994.

[23] P. Milgram and A. F. Kishino, "Taxonomy of Mixed Reality Visual Displays", IEICE Transactions on Information and Systems, pp. 1321–29, 2013.

[24] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J.C. Hart, "The CAVE: Audio Visual Experience Automatic Virtual Environment", Communications of the ACM 35, vol. 6, pp. 64–72, 1992.

[25] H. Rheingold, "Virtual reality". New York: Summit Books, 1991.

[26] MSEAB Weibull, "http://www.mseab.se/The-Virtual-Theatre.htm".

[27] D. Ewert, K. Schuster, D. Johansson, D. Schilberg, and S. Jeschke, "Intensifying learner's experience by incorporating the virtual theatre into engineering education", Proceedings of the 2013 IEEE Global Engineering Education Conf. (EDUCON), 2013.

[28] M. Hoffmann, K. Schuster, D. Schilberg, and S. Jeschke, "Bridging the Gap between Students and Laboratory Experiments". In: Shumaker, R. (Publ.): Virtual, Augmented and Mixed Reality. 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014 : proceedings. Cham, 2014.

[29] K. Schuster, M. Hoffmann, U. Bach, A. Richert, and S. Jeschke, "Diving in? How Users Experience Virtual Environments Using the Virtual Theatre". In: Shumaker, R. (Publ.): Virtual, Augmented and Mixed Reality. 6th International Conference, VAMR 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014 : proceedings. Cham, 2014.

[30] A. de Souza e Silva and D.M. Sutko, "Digital cityscapes". Merging digital and urban playspaces v. 57. New York: Peter Lang, 2009.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

The Design of Interactive Assessment-Cognitive Schema-based System: An Exploratory Study in E-learning Implementation

Melvin A. Ballera, maballera@yahoo.com.ph
Omer Masoud Salih, omeryomery@yahoo.com
Computer Science – Faculty of Science
Sirte University – Sirte, Libya

ABSTRACT

An e-learning website is not sufficient to fully attain the results of online education. There also is a need to align the educational objectives into the design of the assessment to improve and develop cognition, critical thinking and problem-solving skills. Previous studies have explored the potentials of the assessment models but few ventured into their implementation. Others only proposed and introduced conceptual frameworks. The implementation of these proposals, however, revealed that the question type in the assessment phase neglected to align their questionnaire formats into a cognitive schema. At present, the standard multiple-choice question is the most frequently used of the question type of e-learning assessments. However, if this type is the only format adopted by e-learning developers, then the potentially rich and embedded assessment of the computer platform can be will be given up. This paper then focuses on the design of assessment questions, which is created and guided by the hierarchical Bloom cognitive taxonomy and by utilizing rich media formats. Preliminary results conducted for four weeks show a dramatic increase in the academic performance of the students.

KEYWORDS

Bloom cognitive taxonomy, assessments, e-learning, cognitive, interactive

1 INTRODUCTION

Assessment is defined as "a device or procedure used for evaluation by obtaining a sample of a learner's behavior in a specified domain and scoring this behavior in a standardized process [1]. It constitutes a vital part of web-based learning instruction. Through assessment, educational strategists can determine how effective their lessons are in teaching students the intended facts and skills. To effectively assess students, educational strategists must not simply relegate assessment at the end of the learning process or training. This must be also fully integrated into the process of educating students [2]. Assessment designs can greatly influence the learning of the students. It can also be a tool for data gathering and the results gathered can help teachers decide on the performance of the students [3]. At present, many learning e-learning assessments used the standard multiple-choice questions. However, it can be argued that if e-learning developers adapt only this type of assessment, then the potentially rich and embedded assessment of the computer platform will not be totally utilized [4].

Today, the question type currently dominating many e-learning assessments is the standard multiple-choice question. It is necessary for assessment practices to reflect the combinations of acquired skills and knowledge. The complexity and use of these combinations will enable students to interpret, analyze, evaluate problems and explain their arguments. These assessments, which should be fully integrated into the learning process, provide information about the learner's progress and support them in selecting appropriate learning tasks. The consistency of the content, methodology, and

ISBN: 978-1-941968-02-4 ©2014 SDIWC

the manner of assessment will make teaching become more effective. Therefore, it is a worthy undertaking to invest in the design of performance assessments because assessment provides multidimensional feedback for fostering learning [5].

The objective of this paper is to present an assessment questionnaire format by adapting a number of the assessment designs which were investigated and discussed in the related literature. These designs are redesigned and realigned into the Bloom Cognitive Theory Schema and is presented in a more interactive way to suit the computer science curriculum at tertiary level. The paper is organized according to the discussion of related literature, methodologies, initial findings and lastly, the conclusion and future works.

2 RELATED LITERATURE

For the alignment of an effective assessment design, three components are investigated in this section: 1. the search and adaptation of the existing assessment design that can be useful in the computer science curriculum; 2. The design of the assessment according to Bloom Cognitive Taxonomy schema (manner or way to present question in accordance to cognitive prescription) and 3. The incorporation of simulations interactivity in the assessment process.

2.1 E-Learning Assessment Questions

There are several ways to make assessment items innovative and creative. The use of technological enhancements of sound, graphics, animation, video or the incorporation of media can be utilized also for e-learning assessment designs [6]. Figure 1 below shows the summary of the 13 questions types collected from 15 various sources of scientific research and publications. Each question type has a different cognitive level and requires demonstration of varying skills from the learners during assessment process. Although there are many existing assessments that used various question types, few have been tested from the computer science perspective and at the level of tertiary education. Majority of these assessments were implemented in a pencil-paper test and few transformed these assessments into digital form [7], [8], [9], [10], [11], [12], [13], [14]. If such assessment features will be implemented fully into the e-learning, the system will hypothetically deliver cognitive gain among students.

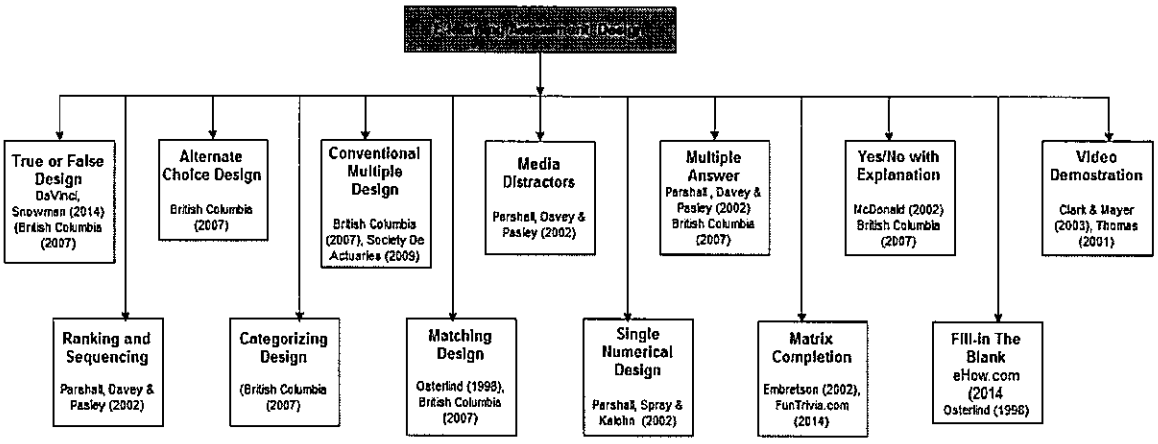


Figure 1: Question Types for E-Learning Platform

2.2 Bloom Cognitive Schema

The cognitive domain of Bloom involves knowledge and the development of intellectual skills, therefore it is necessary to align assessment according to this schema. This schema includes recall or recognition of specific facts, procedural patterns and concepts that help in the development of intellectual abilities and skills [15]. There are six major categories, starting from simplest behavior to the most complex in this schema. The categories can be viewed as degrees of difficulties [16]. Layer one is, "Remembering" which entails establishing definitions, creating fact charts, lists or oral activities. Layer two, "Understanding", includes producing drawings or summaries." Applying" is layer three, and models, presentations, interviews or simulations are applied to new situations. Analyzing" is layer four which includes "distinguishing" between the parts creating spreadsheets, surveys, charts, or diagrams. Evaluating, which is layer five involves critiquing, recommending, and reporting. Putting the parts together in a novel and unique way falls in the sixth layer which is Creating [17]. At present, this model becomes a basis in developing e-learning by transforming its contents, instructional delivery and most importantly the assessment. The layers represent the levels of learning and increasing complexity.

Figure 2 shows the cognitive levels in Bloom's original taxonomy, arranged in ascending order. Each step suggests activities for the specific level. A list of verbs which are commonly used to create learning objectives can be found below each step. When Bloom created this cognitive schema, he intended to use this in assessing the expertise in order to develop new ways in measuring what college students learn. At present, this model becomes a basis in developing e-learning by transforming its contents, instructional delivery and assessment to suit the learners' needs. His work contributed greatly in shifting the focus of educators to learning from teaching.

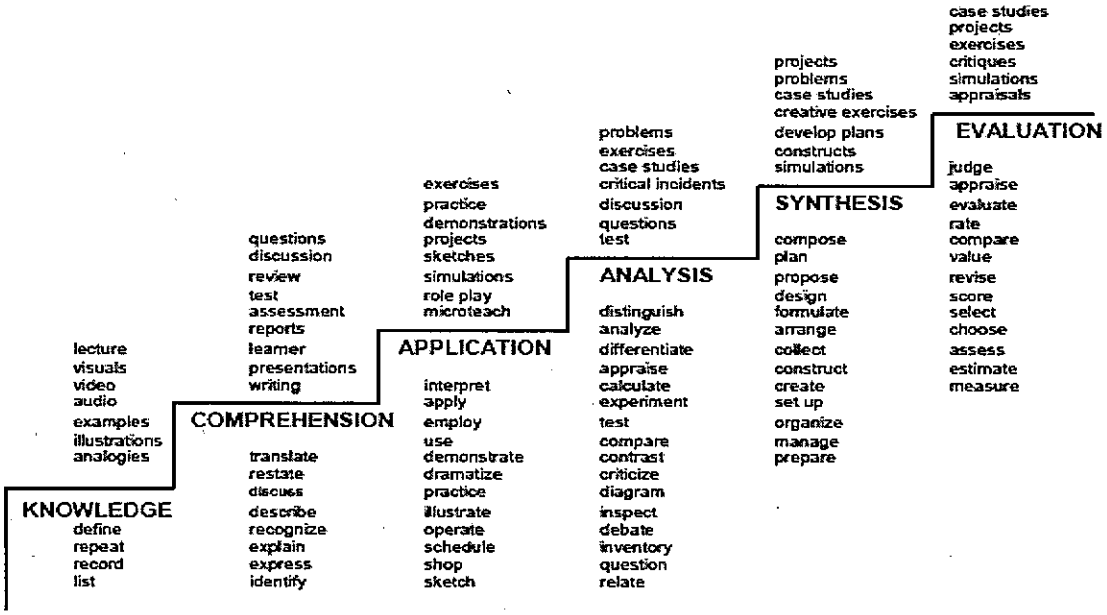


Figure 2: Bloom's Taxonomy Staircase [17]

ISBN: 978-1-941968-02-4 ©2014 SDIWC

2.3 Multimedia and Interactivity

Many educators believe that interactive e-learning assessments allows "learning by doing", arouses interest and generates motivation. Interactivity leads to a more meaningful learning because students are able to test their comprehension, learn from their errors and make sense of what is unpredictable. It can also improve the students' knowledge and performance during the assessment process [18]. Simulations and modelling tools are the best examples of complex, meaningful interactivity in assessments. Such applications models represents a real or theoretical system, and allows users to manipulate input variables, change the system's behavior and view the results. With such applications, learners can construct and receive feedback as a result of their actions. Inclusion of interactive simulations in e-learning assessment improves the quality and outcomes of e-learning. Simulations and visualization tools make it possible for students to bridge experience and abstraction which help to deepen their understanding of ambiguous or challenging content [19]. Interactivity, when used in assessment, is a factor that has the biggest impact on cognitive learning and is the most powerful model of instruction [20].

The use of multimedia in assessment such as graphics refers to the variety of illustrations that include line drawings, charts, photographs, motion graphics such as animation and video. These multimedia can indeed increase learning. Research shows that graphics improve learning through cognitive exercises, storing and retrieving ideas. Mayer claims that a student who practices on assessment with text and graphics is claimed to gain an average of 89% on transfer text as compared to those students who rely on texts alone [21]. It is also found out that the integration of text near the visuals during assessment yielded an average improvement of 68%. Furthermore, explaining graphics with audios followed by a question improved learning t by almost 80%.

Adapting question types from different researches and re-aligning its questionnaires or small tests into cognitive model and presenting it in an interactive and simulative manner can thus hypothetically guarantee learning.

3 METHODOLOGY

3.1 Respondents

The study is organized within the context of Design and Analysis of Algorithms class which is taught at Sirte University, Libya. The entire data collection and training lasted for 4 weeks for initial testing. All students are familiar with the use of electronic materials and had seen the implementation of the e-learning system and were given one week familiarization of the system flow and navigation. During the training, student were given several examinations (diagnostic, formative and summative) to determine their knowledge level of the course. The passing mark is 75.

Prior to implementation, students were informed about the research and the task involved. Students were given time to navigate the e-learning system so that they would be familiar and be directly involved in the learning process. If issues arouse during the learning process, the researcher provided necessary assistance in support for blended learning. At the beginning, participants were given diagnostic assessments, while at the end of each lesson, a formative assessment is given. All students were subjected to summative assessment at the end of the course training.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

3.2 Data Collection

In this study, primary data were collected in two ways. The first was the experimental collection where various tables were populated dynamically, manipulated and extracted to generate several reports such as examination results, graphs, frequency of the practice examination and trials. The second was the survey which was divided into two parts. The first part was the measure of the internal reliability of all the questionnaires stored in the Item Bank. The second part was the acceptability of assessment design factors. Factors that affect the assessment design were content of the item, the visual design (colors, balance, readability), accessibility (links, feedback and explanation facilities), assessment types (difficulty, bloom level), navigation (transition of questionnaires, pop-up windows, reminders), learning support (specific part of the lesson, additional references) and interactivity. To measure the internal consistency of the questionnaires, the Cronbach alpha was used while z-test was used to evaluate the acceptability of the assessment design factors.

Degree of Difficulty

The 13 question types investigated and presented in the literature were categorized according to the Cognitive Bloom Taxonomy. Table 1 shows the question types description and the degree of difficulty *df*, for each type in different assessment formats. In formative assessment, the *df* is 1 for reviewing purposes and practice examination at the end of each lesson. The *df* of Bloom Cognitive examination (diagnostic) on the other hand is also 1, to measure the cognitive improvements of the learner which is usually administered every two weeks of the training.

The *df* of summative assessment differs accordingly since it is the most important performance matrix. As the Bloom category goes to the bottom of the table, the more difficult the questions and the deeper the cognitive development become. Every question has a level of difficulty, and this level is also utilized upgrading the students' performance matrix. Higher ability is demonstrated when a student answers harder questions that correctly answering an easier question. The *Remember* category has *df* 1 while the *Understand*, *Application*, and *Analyze* categories have a *df* of 1.5 while *Evaluate* and *Create* have *df* of 2.

Table 1: Questions Types and their Degree of Difficulty (*df*)

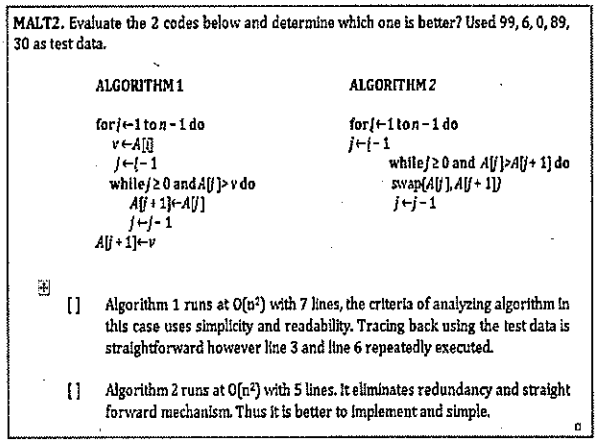
			Degree of Difficulty		
Bloom Taxonomy	Question Types	Description	Summative	Formative	Diagnostic
REMEMBER	MATF	Multiple True or False Questions	1	1	1.5
	MTCQ	Matching abd Categorizing Questions	1	1	1.5
	TOFQ	True or False Question	1	1	1.5
UNDERSTAND	MCMA	Multiple Choice Multiple Answer	1	1	1.5
	MCID	Multiple Choice with Illustrative Diagram	1	1	1.5
APPLICATION	CSMA	Complex Single Multiple Choice Questions	1.5	1	1.5
	SNCQ	Single Numerical Construction Questions	1.5	1	1.5
ANALYZE	SMCQ	Situational Multiple Choice Questions	1.5	1	1.5
	SAMC	Single Answer with Enumeration Questions	1.5	1	1.5
EVALUATE	MSOQ	Matrix Completion Questions	2	1	1.5
	MALT	Multiple Alternative Questons	2	1	1.5
CREATE	FIBE	Situational Fill-in the Blanks and Enemuration	2	1	1.5
	DSVQ	Video Simulation with Audio Play Qiestions	2	1	1.5

ISBN: 978-1-941968-02-4 ©2014 SDIWC

After aligning the colleceted question types as shown in Table 1, questionnaires underwent fornatting using the guided cognitive verbs schema as presented Figure 2 and then the interactivity and simulation to the question were added. The use of graphs, videos and other media formats, and required-response questions was incorporated in the system. The Item Bank is currently a repository of different questions types with varying difficulty level. It contains 280 questions with explanation facilities divided among thirteen (13) questions types and were used to produce the Bloom Cognitive Taxonomy examination, the random formative examination and the random summative examination.

3.4 Question Item Design and Interactivity

For brevity, two live illustrative question types were extracted from the system prototype for the purpose of illustration. Alternate choice items are somewhat similar to true/false type of questions. However, rather than letting the students determine whether a single statement is correct or not, this type of questions asks the student to select the better answer between two choices. Choices are often scenarios or cases, as shown in Figure 3 below.


MALT2. Evaluate the 2 codes below and determine which one is better? Used 99, 6, 0, 89, 30 as test data.

ALGORITHM 1

```
for i←1 to n-1 do
    v←A[i]
    j←i-1
    while j≥0 and A[j]>v do
        A[j+1]←A[j]
        j←j-1
A[j+1]←v
```

ALGORITHM 2

```
for i←1 to n-1 do
j←i-1
        while j≥0 and A[j]>A[j+1] do
        swap(A[j], A[j+1])
        j←j-1
```

[ ] Algorithm 1 runs at O(n²) with 7 lines, the criteria of analyzing algorithm in this case uses simplicity and readability. Tracing back using the test data is straightforward however line 3 and line 6 repeatedly executed.

[ ] Algorithm 2 runs at O(n²) with 5 lines. It eliminates redundancy and straight forward mechanism. Thus it is better to implement and simple.

Figure 3: Alternate Choice Example

In this type, students were shown two possible algorithmic models for computing their running time complexity and must choose the most accurate response option. In this case, the correct answer was the second option due to its simplicity. Innovations in the multiple-choice category for online settings can include new response actions not commonly found in paper-and-pen settings. This entails clicking on an area of a graphical image. It can also include new media, such as sound clips which can be considered as distractors. Such new media innovations are represented in Taxonomy as Multiple Choice with Illustrative Diagrams. An example is given in Figure 4.

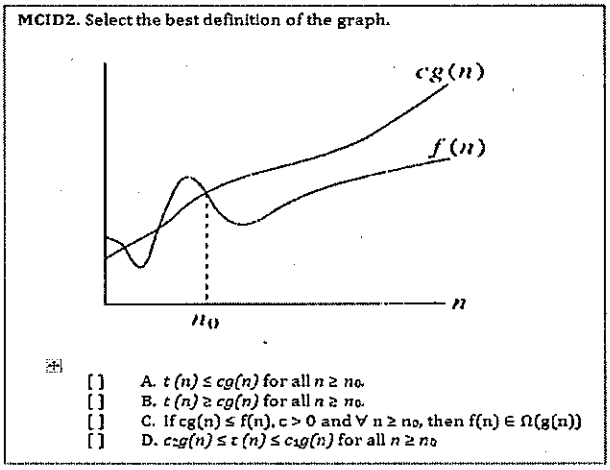


Figure 4: Multiple Choice with Illustrative Diagrams Example

In this example, respondents must select one of the four choices that corresponded to the meaning of the graph. There were four choices to choose from. This is similar to the standard multiple choice question but aside from choosing from the four possible answers, this method of response involves also analysis.

Many interactive activities were included in the assessment design to give learners the "personal touch and control" in the assessment process. Student could write their answer using the fill-in the blank question types, compute the next sequence and analyze the pattern in completion matrix question type. Students could also enumerate answers, view and analyze graphs

and allow feedback. The explanation facilities could also derive the solution and link student's misconception into specific part of the learning materials. To enhance the learning process further, videos, and other simulative process were incorporated into the system to allow the method of "learning by doing". Student could view the algorithm and its simulative effects given certain inputs and variables.

4 RESULTS AND DISCUSSION

4.1 Internal Consistency and Z-test Results

Prior to the post survey for students, the survey forms were presented among the academic staff to validate the measurement scale and questionnaires. The Cronbach's Alpha coefficient for internal consistency reliability test was used for each scale. Cronbach's alpha reliability coefficient normally ranges between 0 and 1. It provides the following rules of thumb: $\alpha \geq .9$ – Excellent, $.7 \leq \alpha < .9$ – Good, $.6 \leq \alpha < .7$ – Acceptable, $.5 \leq \alpha < .6$ – Poor and $\alpha < .5$ – Unacceptable [22]. The results of Cronbach's Alpha coefficients for each scale are presented in Table 2.

The results indicated that all scales satisfied the requirement for internal reliability. All Cronbach's alphas of the scales were higher than .60. The lowest value of Cronbach's alpha is .62 in Accessibility scale while the highest is .74 in Navigation scale. The impact of the reliability of each question in the survey can be determined by calculating Cronbach's alpha the ith variable for each $i \leq k$ is deleted. Thus, for a test with k questions, each score x_j alpha was calculated for x_i for all i where $x_i = \sum_{j \neq i} x_j$.

Table 2: Cronbach's Alpha Coefficient for each Measurement Scale

VARIABLES	CRITERIA						
	Content	Visual Design	Accessibility	Assessment	Navigation	Learning Support	Interactivity
k	5	5	5	5	5	5	5
Cumulative Variance (Vi)	0.86	0.99	1.08	1.00	0.99	1.01	1.01
Variance	1.75	2.06	2.14	2.14	2.41	2.08	2.31
α (Alpha)	0.63	0.65	0.62	0.67	0.74	0.64	0.70

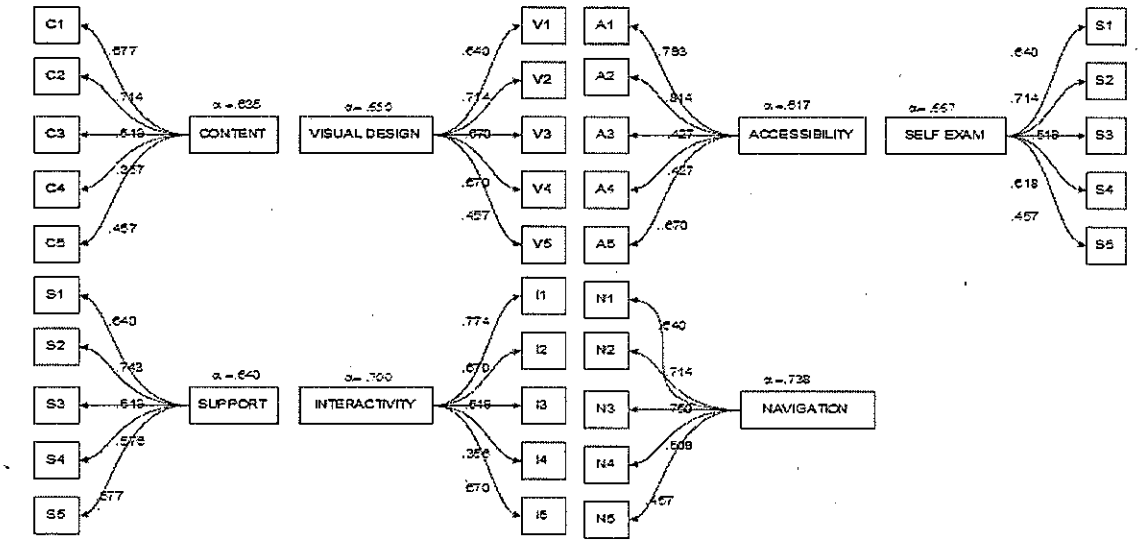


Figure 5: Reliability Coefficient after Deleting an Item

Figure 5 shows that the overall reliability for Content is .636 while individual reliability of questionnaire within the scale are: for C1 is .677, C2 is .774, C3 is .519, C4 is .337 and C5 is .457. In this scale, C4 was the most affected and could be deleted from the survey form. The Visual Design overall reliability is .650 and the most affected was V5 with Cronbach value of .457. On the other hand, Accessibility scale overall reliability is .617 and the most affected questions were A3 and A4 which both have values of .427. Similarly, with the remaining scale, questions with smaller Cronbach alpha compared to the overall scale reliability were the most affected and could be deleted from the survey form. If the reliability coefficient increased after an item was deleted, it can be assumed that the item was not highly correlated with the other items. Additionally, the decrease in reliability coefficient can lead to the assumption that the item is highly correlated with other items. [23]. As shown in the table, the omission of any single question does not affect the Cronbach's alpha very much. Questions with low reliability compared to its overall measurement scale were not deleted because small set of questionnaires affects the reliability value [24]. In this case, five questions in each measurable scale were acceptable and there was no need to delete the item since the uniqueness of each item could easily be seen. According to Cortina [25], the uniqueness of the item can be assessed with the coefficient alpha.

Table 3 shows the results of the post survey conducted among staff members to determine the overall reliability of the software and the 280 questionnaires stored in the Item Bank. These questions were used for various assessments employed in the prototype. During the survey, random questions were shown from the system to evaluate and rate their reliability. The overall internal consistency of the software is .81, which is considered good while the overall reliability for the 60 questionnaires for Bloom Cognitive Taxonomy is .84. Similarly, the internal reliability for questionnaires which were used for formative and summative assessment is .72.

Table 3: Crobach's Reliability of Questionnaires and Overall Acceptability

	Software Acceptability	Bloom Taxonomy	Questionnaires
K	13.00	3.00	3.00
sum var	7.10	2.22	1.13
var	28.64	5.02	2.19
α= alpha	0.81	0.84	0.72

Table 4: Z-test of Different Measurable Scale

Criteria	Mean	Standard Deviation	z
Content	4.37	0.79	2.89
Visual Design	4.29	0.69	2.57
Accessibility	4.26	0.76	2.13
Self Assessment	4.34	0.88	2.40
Navigation	4.21	0.62	2.09
Learning support	4.29	0.93	1.92
Interactivity	4.26	0.92	1.76
Motivation	4.29	0.80	2.22

ISBN: 978-1-941968-02-4 ©2014 SDIWC

An important concept in the evaluation of assessments and questionnaires is the Alpha. It is required from the assessors and researchers that they estimate this quantity to add validity and accuracy to the interpretation of their data. A low value of alpha can mean a low number of questions and poor interrelatedness between items or heterogeneous constructs. For example, if a low alpha is due to poor correlation between items, then some items should be changed or totally eliminated. If an alpha is too high, it may suggest that some items are repetitive as they evaluate the same questions in a different manner [26]. As observed in the study, the overall alpha is not too high but still considered highly acceptable at all levels.

Table 4 shows the summary of the perception of students on the significant level of different assessments scales. The mean is given with its standard deviation. The highest mean is 4.37 from the Content scale while the lowest is 4.21 from Navigation scale. The z-values at $z_{.05}$ = 1:645 makes all the critical values of measurable scale significant using one-tailed critical region. The z-values computed are greater than tabular value at alpha of .05. Based on Likert scale, the mean of each measurable variable is higher than the agreeable level which was successfully correlated by the z-test. Table 5.1 shows a live data extracted from the prototype for 4 weeks using the link http://maballera.byethost7.com/elearning/. For the purpose of illustration, a number of records were selected from the different tables in the database. The table shows that during diagnostic exams, the items correctly answered by students gradually increased. This exam was composed of 30 questions. The questionnaires or items were designed according to the Bloom Taxonomy assessment. The table also reveals that the the number of trial decreases as the weeks of trainings continued. This can be attributed to the familiarity of the students with the assessment structure as they continue doing the process. The number of trials determined the number of time a formative assessment was taken to reach the competency level. For example, formative assessment shown in W1, first row indicates that students had to take this assessment ten times before they obtained a competency score of 7. As observed, only the 6, 7 and 8 scores were recorded in the formative results. Six (6) is the minimum score which is 75% out of 8. The system could load eight random questions during practice exams.

Table 5: Diagnostic, Number of Trials and Formative Assessments

Diagnostic Results				Number of Trials				Formative Results			
W1	W2	W3	W4	T1	T2	T3	T4	W1	W2	W3	W4
5	8	9	11	10	8	6	5	7	8	7	7
7	8	11	12	8	6	5	6	7	7	7	8
8	9	12	14	9	5	4	4	7	7	7	7
8	12	12	16	7	5	4	4	7	8	8	8
9	13	14	21	8	6	5	4	7	8	7	8
9	14	16	22	9	7	4	5	8	8	7	8
7	15	20	25	5	4	6	3	8	8	7	8
6	8	14	17	6	5	7	5	7	8	7	8
8	9	14	18	7	6	4	3	7	8	8	8
6	8	17	19	8	6	4	2	7	8	8	8

ISBN: 978-1-941968-02-4 ©2014 SDIWC

5 CONCLUSION

This paper successfully combined 13 question types extracted from 14 publications. It also aligned the 280 questionnaires stored in the Item Bank according to cognitive schema. The cognitive schema was composed of different "verbs" words which served guide in creating questionnaires that support hierarchical cognitive development. The questionnaires were reproduced as part of the e-learning assessment with added interactivity and simulations. The questionnaires stored in the Item bank were measured using internal reliability test and all were at acceptable level. The design factors of the assessment level were statistically significant at all assessment measurement scale. Based on the preliminary results of the study, students improved their academic performance. The number of trials in taking the practice assessment became less as the results increased. The success of the initial testing was attributed to the design of the assessment which allowed the students to review and reload the questionnaires several times thereby making them familiar with the graded assessment. Being interactive, the item or the question was linked to the explanation facilities, specific learning materials and review module. Although the initial results are quite convincing and acceptable, a thorough study is needed to establish the impact of the design in the diagnostic, formative and summative assessment.

6 REFERENCES

[1] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.. *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association. 1999.

[2] M. Birenbaum, M. New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards.* Dordrecht, The Netherlands: Kluwer. (pp. 13–36). 2003.

[3] K. Scalise & B. Gifford. Computer-Based Assessment in E-Learning: A Framework for Constructing "Intermediate Constraint" Questions and Tasks for Technology Platforms. *Journal of Technology, Learning, and Assessment, 4*(6). 2006.

[4] E. Schreiner.. How to Design Effective Classroom Assessments. Retrieved July, 22, 2014 from http://www.ehow.com/how_7408048_design-classroom-assessments.html

[5] J. Cowan. Designing assessment to enhance student learning. Retrieved June 10, 2014 from http://www.heacademy.ac.uk/assets/ps/documents/practice_guides/practice_guides/

[6] C. G. Parshall, T. Davey & P.J. Pashley. Innovative Item Types for Computerized Testing. In W. Van der Linden, Glas, C. A. W. (Ed.), *Computerized Adaptive Testing: Theory and Practice.* Norwell, MA: Kluwer Academic Publisher. pp. 129–148, 2000.

[7] C. G. Parshall, J. Spray, J. Kalohn, & T. Davey, T. Issues in Innovative Item Types. In *Practical Considerations in Computer-Based Testing.* New York: Springer. pp. 70–91, 2002.

[8] S. Embretson. Generating Abstract Reasoning Items with Cognitive Theory. In S. Irvine, Kyllonen, P. (Ed.), *Item Generation for Test Development.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers. pp. 219–250, 2002.

[9] M.E. McDonald. Developing Multiple-Choice Items. In *Systematic Assessment of Learning Outcomes.* pp. 83–120, 2002.

[10] British Columbia. Examination Booklet. Retrieved July 7, 1014 from http://www.bced.gov.bc.ca/exams/search/grade11/english/sample/exam/0708ss_P.pdf

[1] S. J. Osterlind. *Constructing Test Items: Multiple-Choice, Constructed-Response, Performance, and Other Formats.* Norwell, MA: Kluwer Academic Publisher. 1998.

[12] Society De Actuaries. Construction and Evaluation of Actuarial Models. Retrieved July 8, 2014 from

http://www.soa.org/files/pdf/edu-2009-fall-exam-c-questions.pdf

[13] FunTrivia.com. True or False Trivia. Retrieved August 21, 2014 from http://www.funtrivia.com/quizzes/general/true_or_false.html

[14] EHow.com.. How to Write a Fill In the Blank Questions. Retrieved August 21, 2014 from http://www.ehow.com/how_8233244_write-fill-blank-questions.html

[15] Bloom, B. Mastery learning. New York: Holt, Rinehart, & Winston. 1971.

[16] L. W. Anderson, & D. R. Krathwohl. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. Allyn & Bacon. Boston, MA (Pearson Education Group). 2001. http://epltt.coe.uga.edu/index.php?title=Bloom's_Taxonomy

[17] A. Churches. Bloom's Taxonomy Blooms Digitally. 2008. http://www.techlearning.com/studies-in-ed-tech/0020/blooms-taxonomy-blooms-digitally/44988

[18] M. Rosenberg, M. E-Learning: Strategies for delivering knowledge in the digital. 2000.

[19] R. Thomas.. Interactivity and Simulations. Multi-verse Solutions Ltd. 2001. http://www.jelsim.org/resources/whitepaper.pdf.

[20] R. E. Clark. & T. G. Craig. "Research and Theory on Multimedia Learning Effects." In: M. Giardina (red). "Interactive Multimedia Learning Environments. Human factors and technical considerations on design issues.. NATO ASI Series. s. 19-30. 1992

[21] R. Mayer. The promise of multimedia learning: using the same instructional design methods across different media. Learning and Instruction. Vol. 13. Pp. 125-139. 2003

[22] D. George & P. Mallery,. SPSS for Windows step by step: A simple guide and reference. 11.0 update (4th ed.). Boston: Allyn & Bacon. 2003.

[23] C. Zaiontz. Real Statistics Using Excel: Cronbach Alpha. Retrieved January 21, 2013 from http://www.real-statistics.com/author/

[24] M. Tavakol & R. Dennick. Making sense of Cronbach's alpha. *International Journal of Medical Education*. 2011; 2:53-55 Retrieved June 12, 2014 from http://www.ijme.net/archive/2/cronbachs-alpha.pdf

[25] J. M. Cortina. What is coefficient alpha? An examination of theory and applications" *Journal of Applied Psychology. pp.* 98–104. http://psychweb.psy.umt.edu/denis/datadecision/front/cortina_alpha.pdf

[26] D. Streiner. Starting at the beginning: an introduction to coefficient alpha and internal consistency. Journal of personality assessment. 80:99-103. 2003.

Coordinating Mobile Servers for Static Hierarchical States

Savio S.H. Tse[1] and Markus Schaal[2]
[1]Computer Engineering Department, Istanbul University
Avcilar Campus, Avcilar, 34320 Istanbul, Turkey.
sshtse@gmail.com
[2]Webtrekk Gmbh, Berlin, Germany
schaal@acm.org

ABSTRACT

We design a Peer-to-Peer network to maintain a large set of hierarchical static states. We argue that these states are common and natural in collaborative knowledge-based systems, and online games. On the top of the hierarchy, we apply many B+-trees of order-k for connecting all online nodes to enhance parallelism, where k is any constant more than two. The overhead communication cost for each join and leave is bounded by $O(\log_k N)$ messages, and the number of connections (edges) in each node is bounded by 2k+4.

KEYWORDS

Peer-to-Peer network, dynamic network, collaborative computing, static hierarchy.

1 INTRODUCTION

Peer-to-Peer (P2P) networks is a popular research topic nowadays. Due to the page limit, we skip the discussion on its general features. We agree with the definition in [11], which says that inviting users to become servers is basically the idea of P2P networks. In this paper, we use a P2P network to coordinate mobile servers for maintaining a collection of static hierarchical states. By static, we mean that the hierarchy of the state structure will rarely change, while the attributes of states can change often. There are quite a few examples of unlimited numbers of static hierarchical states. The most obvious one is knowledge about the world, which is dispersed in time and space at the physical level and becomes more and more abstract at higher and higher levels of the aggregating hierarchy. Of course, many concrete scenarios for the world knowledge hierarchy can be thought of. Another example is online game. Knutsson *et al* gave a P2P system for supporting multiplayer games [6]. Though it is not the main stream of research in P2P network, there are more and more works on this interdisciplinary area [1], [4]. Inside an online game, there can be infinitely many static objects, which fall into a hierarchy. For example, a chimney belongs to a house, and the house belongs to a region. The last scenario for static hierarchy is an information system for emergency management exploiting collective intelligence or crowd sourcing. For example, the temperature outside our windows can be very useful for regional weather forecasting.

To support a large static hierarchy, we design a P2P network that consists of a regular root manager plus many mobile servers. Ideally, P2P nodes should be organized in a balanced tree which matches the state hierarchy. However, due to dynamic infrastructures, where nodes can join and leave freely, the tree cannot always be balanced, and in the worst case, it can become a forest of disjoint trees. Hence, we propose an auxiliary structure and algorithms for join and leave in this work that bound the performance overhead as proportional to the logarithm of the number of existing P2P nodes only. It is well-known that the minimum network diameter decreases

ISBN: 978-1-941968-02-4

with the maximum number of edges of a node. However, we cannot over-burden a node by too many edges. The *limited scale-free* network is an example where a limit is applied to the degree of a node [2]. For facilitating dynamic networks, we further push the limit to 2k+4, where k is a small constant, in this paper. We apply a k-ary tree for modeling the static hierarchy, and we connect the online nodes by a number of B+-trees of order k. The states in the hierarchy are uniquely assigned to members when they register in the system. When a member is offline, his duty to maintain the state will be handled by another member. When he returns, he takes back his duty. Along with this, each member will have a unique identity.

2 RELATED WORK

There have been many topologies developed for P2P networks. In Gnutella [5], since the network features are decentralization, anonymity, and autonomy, it is difficult to even perform the fundamental operation---searching---efficiently. The searching method is a simple flooding, and this incurs high communication cost. In Freenet [3], Clarke *et al* applied a DFS-like searching technique with limited number of hops. It saves many messages comparing with flooding but faces the problem of unreachability. The Content Addressable Network (CAN) model, in [8], maps the nodes onto a multi-dimensional Cartesian coordinate space. Precisely, each (key,value) pair is deterministically mapped to a point in the space by a uniform hash function. Tapestry [15] has a hierarchical overlay structure that gives a quick convergence from different sources to any single destination. Pastry [10] uses the nodeID to indicate its position such that routing to the numerically closest node to a given key in fewer than $(1/4)\lceil \log N \rceil$ steps. The expected number of steps is O(log N). Chord [12] employs a cyclic name space like Pastry. Unlike Pastry, its routing is confined to its logical ring structure. Chord's mechanism is also employed in [7]. These examples share the commonality that node identifiers and infrastructures are not arbitrarily chosen, but designed to facilitate searching.

In this paper, we consider very large P2P networks. A source node makes use of the P2P hierarchy to send an invitation to the destination node with which it wants to communicate. After they exchange their sockets---network addresses and port numbers, they do not need the P2P hierarchy for further communication. This method frees the hierarchy from heavy traffic between peers.

We expect the sizes of P2P networks to rocket as the technology becomes more sophisticated and popular. We hope that P2P networks will no longer be limited by file sharing. However, while the research on P2P networks has been very hot for years, it is observable that the development on applications has remained still for years. Walkerdine *et al* showed the same observation in [13]. One reason may be due to the lack of a design framework [13]. The actual reasons may be complicated, and one can be the lack of a rewarding scheme for the serving peers, because peers need more incentive to take more responsibility in the system, and the application will then have vitality.

3 STATIC STATE HIERARCHY

Let k be a positive integer that is more than two. The state hierarchy is basically a logical k-ary tree. Each parent has at most k children. In the hierarchy, one edge is connected to each leaf state, at most k edges are connected to the root state R, and at most k+1 edges to each internal state. We can represent the identities in dot format. R has an identity (0), and its children have (0.1), (0.2), ..., and (0.k), respectively. The children of (0.i) are (0.i.1), (0.i.2), ..., (0.i.k), where $i \leq k$. Each identity can also be viewed as a base-k number and has a unique numerical value. Two identities with

ISBN: 978-1-941968-02-4 ©2014 SDIWC

different lengths can be ordered according to numerical values after appending zeroes to the right of the shorter one. It is easily seen that the leftmost digit can be omitted, however, we keep it for ease of discussion.

The k-ary tree may not a full tree, as the parent-children relationship depends on the natures of the objects concerned. For example, the state of a house is the parent of the state of a window inside. If a parent x, by nature, has more than k children, detach them from x, split them into a minimum number of groups of sizes at most k, and for each group, create a parent, and attach it to x as a child. These steps are repeated until x has at most k children.

4 PHYSICAL NETWORK MODEL

The network in question is modeled by an arbitrary N-node undirected and unweighted connected bounded degree graph G=(V,E), where $V=\{v_1, v_2, \ldots, v_N\}$ is the set of nodes, and E the set of edges. The degree of each node is at most 2k+4. V refers to the set of online nodes in the system, and hence, it is dynamic. The P2P network overlays a transport layer and network layer protocols. Each node knows its own network address and transport layer port number. We assume that they are TCP and IP, and simply refer to them as one *physical address*. A point-to-point connection (an edge) in the P2P network requires a TCP connection, which costs at least three messages to establish the TCP plus some messages to connect the P2P application layer. A P2P hop can induce many hops in network layer, and therefore only the messages for exchanging physical addresses route along the P2P network edges. Further communication is done through the connection by the nodes' physical addresses.

A node is engaged to a state when it first applies to be a member. The member takes the state identity as its own. This engagement is permanent, even when the member is offline, until it terminates its membership. Hereafter, a node is referred to a member, a member is called an existing node, or online node, if it is online; otherwise, it is called a departed node, or offline node. The system manager is engaged to the root state R. It is always online. It can be a single-node host, as well as a cluster of hosts. For simplicity, we call the manager R, and let it be a single-node host because it makes no difference to the algorithms in this paper. At the beginning, when R is the only member, it is responsible for all states in the system. When a new member registers, R will assign an un-engaged state to it. It will then be responsible for all un-engaged descendant states. The member is also responsible to (but not engaged to) states that are its nearest engaged ancestor. These two duties constitute the *primary duty* of the new member, and is changeable if some ancestor or descendant states are engaged to other new members. A member is free to be online and offline. In order to maintain the robustness of the system, we assume that the existing nodes are all robust and that they can leave only in a "polite" way, such that enough information exchange is completed before departure. (Without this assumption, regular probing and replication are sufficient for remedial actions.) We assume the P2P network is asynchronous. A task's performance is measured by its message complexity, while each message is referred to the one in the P2P layer. It is the cost to the whole network. The cost to each node is measured by its connection complexity, which is the number of edges it has.

We call the state hierarchy the *state tree*. Every node knows the physical address of each of its online parent and children, if any. Note that the state tree may not necessarily be a physical one, as the members, except R, can be offline, and some states may not be engaged. Even if the state tree is connected, its height can be O(N) in the worst case. Clearly, this deficiency is due to the fact that the state hierarchy is static. Moreover, when a member

ISBN: 978-1-941968-02-4 ©2014 SDIWC

is offline, the states it is responsible for are no longer accessible. In order to tackle these problems without changing the static structure, like the approach in [14], we construct an adaptive connected structure, as explained below.

5 IMAGE AND SUBIMAGE NODES

We first define some terminologies. In the state tree, a *leaf node* is an online node that has no online descendant; and a *junction node* is a node that has at least two children, with each of them having a descendant leaf node. A junction node can be online or offline. Clearly, the sets of these nodes are dynamic.

On top of the primary duty, a node may have up to two kinds of *temporary duties*. A temporary duty is due to other nodes leaving and it can be released or reduced when the corresponding nodes return. The first kind of temporary duty is to store and maintain states' information. When a node is leaving the system, the responsibility of storing and maintaining states will be handed over to its nearest online ancestor. It is easy for each online node to search for its nearest online ancestor; we will discuss how at the end of Section 6. Although this allocation of responsibilities for states is temporary, it is consistent with members' initial assignments. The second kind of temporary duty is to propagate invitations on behalf of the corresponding departed node. However, if this duty is again passed to its nearest online ancestor, the ancestor can be burdened by too many connections if the other descendants request the same. Therefore, we share this duty with *image nodes*. The image node of a node is chosen from one of its descendant leaves. An internal node has exactly one image node, but an image node can serve at most one junction node plus many other nodes. When an internal node is in the system, its image node exists but does not perform any task for it. However, when the internal node is about to disconnect from the system, it invokes its image to take up all its edges in the state tree. The duty of an image node is to propagate invitations through these edges. Recall that a leaf can be the image node of some internal nodes, which must be its ancestors, but it itself does not need any image node because it is a leaf and need not propagate invitations up or down in the state tree. Three rules for the image scheme follow:

Rule 1 One internal node has exactly one leaf as its image, and they are connected by at most $2\log_k N$ hops.

Rule 2 If a leaf l is the image node of $p \geq 1$ internal nodes, then (a) all these nodes are ancestors of l; (b) they are in two disjoint blocks of sizes p_1, $p_2 \geq 0$, where $p_1 + p_2 = p$; (c) (contiguity) for each block B, the nodes between the nearest and farthest ancestors are all in B; (d) (locality of the lower block) if p_1, $p_2 \neq 0$, the block nearer to l contains l's parent; and (e) (simplicity) if there is no junction node among these p internal nodes, the farther block is empty.

Intuitively, the ancestors fall into at most two disjoint vertical lines, and the lowest one is connected directly to the leaf node.

Rule 3 If a leaf is the image node of $p \geq 1$ internal nodes, then there is at most one junction node among them; if there are two contiguous blocks, then the junction node should appear in the upper one.

Rule 3 will be used to argue that when a leaf assumes duty as an image node, it will have at most 2k+4 edges, no matter how many internal nodes assigned to it have departed. Rule 1 guarantees that the maintenance cost of the image nodes is bounded by O(log N) messages. Rule 2 is for the coverage of all internal nodes as well as for keeping the bound on the number of edges leaf nodes have.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Consider the case that only one internal node is assigned to a leaf. Rule 1 can be kept true by using an edge between the internal node and its image. Hence, an internal node has no more than (k+1)+1=k+2 edges so far, where k+1 edges are its original connections in the state tree, and the last edge connects to its image. Together with Rule 3, the degree of a leaf node is at most (k+1)+1=k+2 so far, where k+1 edges are from the leaving node, and the last edge is connected to its parent.

Consider the case that there are p>1 internal nodes assigned to an image. Let A be the set of these nodes, and A_1, $A_2 \subset A$ be the upper and lower blocks, respectively, where $A= A_1 \cup A_2$ and $A_1 \cap A_2 = \varnothing$. If there is only one block, we simply assign $\varnothing$ to the corresponding subset. There are at most p-1=O(N) edges in A that should be taken up by the image.

Whenever there are as many as O(p) nodes in A staying in the system and the rest of O(p) nodes have departed (Figure 1), these edges will make its degree exceed the bound 2k+4.

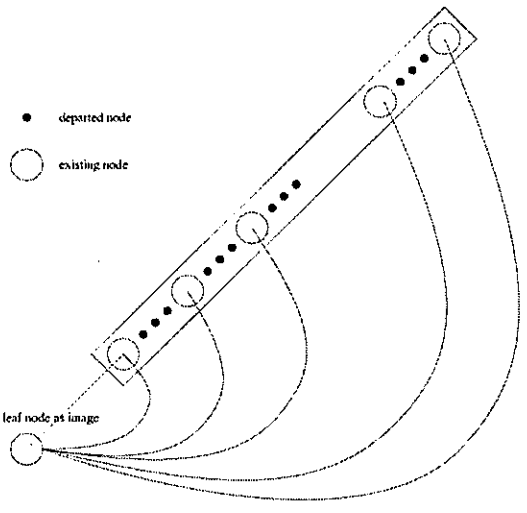


Figure 1. Edges required for a block of internal nodes.

To tackle this problem, we build a B+-tree of order k for each of A_1 and A_2, and the internal nodes inside these trees are called *subimages*, as it helps the image to look after the nodes inside the subsets. We call this tree *subimage tree*. The junction node, if $A_2 \neq \varnothing$, will be directly connected by the image, and therefore, it is not in the tree. The internal nodes, as well as the root, are "simulated" by the bottom nodes such that the root of each subtree is "simulated" by the rightmost descendant of the leftmost child. The root is then connected directly to the image. At the bottom level, nodes are connected by edges in both directions. This prevents a single hop resulting in $2\log_k N$ steps for propagating invitations. The maintenance cost is bounded by O(log N) messages. In the B+-tree, there are at most k+1 edges for a subimage node, plus two more edges along the bottom level. Therefore, every node in A has at most k+1+2=k+3 edges. The image needs only one edge pointing to the root, and an edge connecting to a junction node. Hence, it can have at most 2 extra edges so far. For keeping Rule 1 true, it suffices to apply the subimage trees. The robustness of the subimage tree and Rules 2 and 3 will be guaranteed by the algorithm in Section 6.

A leaf node is called a *free leaf* if it is not an image of any junction node. To provide a fast search for a free leaf in the algorithm described later, a B+-tree of order k is used to store all free leaves. We call it a *free-leaf tree*. Each free leaf needs k+1 more edges for this B+-tree, plus 2 edges at the bottom level. Together with the edge in the state tree, the number of edges for a free leaf is bounded by k+4. Since a leaf node can have no internal node assigned, and a leaf node may become an internal node (and vice versa) after a number of leave and join operations, the relation between internal nodes and their images are dynamic.

The last B+-tree of order k is needed for the junction nodes. We call it the *junction tree*. If

a junction node is online, it then has (k+1)+(k+1)+1=2k+3 edges, where the first term is for the state tree, the second term is for the junction tree, and the last one connects to its image node. If a junction node is offline, its image will be its representative that is in the tree. Then, the image has at most (2k+2)+2=2k+4 edges, where the first term is for the offline junction node, one is for the root in the subimage tree, and one is for the state tree.

6 THE ALGORITHM FOR JOIN AND LEAVE

We show that the three rules can be kept true in O(log N) messages when a node is leaving or returns from being offline.

Suppose that a node x is back in the system. Search the leaf tree and if more than one leaf has a proper prefix equal to x, then either x is a junction node or there is another junction node under x. In the latter case, x belongs to the same block as its immediate upper or lower junction node. Hence, in both cases, search from the junction tree to find the node's image. After x contacts its image and recovers its state tree edges, it performs the maintenance on the subimage trees and then has fulfilled its duty.

If only one leaf has a proper prefix equal to x, then x is not a junction node, and the leaf is its image. Same as above, x needs to get back its edges, and perform the maintenance of the subimage trees.

If no leaf has a proper prefix equal to x, then x will become a leaf. Find the nearest leaf y---the one that has the longest common prefix with x. Let a be the node whose identity equals the longest common prefix of the identities of x and y. Obviously, a is the nearest ancestor of x such that a or its image must be in the system. Note that a can be y. a or its image can be found by the junction and subimage trees. Connect x to a or its image by an edge if a is the parent of x. If there are some departed nodes between x and a, then x will be the image of the departed nodes and an edge is still needed between x and a or its image, as x takes all edges for the departed nodes. The maintenance of the leaf tree, and the free leaf tree should be done for the insertion of x. We have three cases for a:

Case 1: a was a leaf and becomes an internal node.

Consider Rule 2. Suppose that a is the image of two blocks of internal nodes; otherwise, we can assign any of them to be empty. x takes the upper block of internal nodes of a, if any, without any changes. The lower block of x is the union of the lower block of a and the set storing the nodes between a and x, including a. Rule 2 is kept true, and obviously, Rule 3 is also true as the only possible junction nodes are always in the upper block. Rule 1 is kept true after the maintenance of the subimage tree is performed. Figure 2 shows an example.

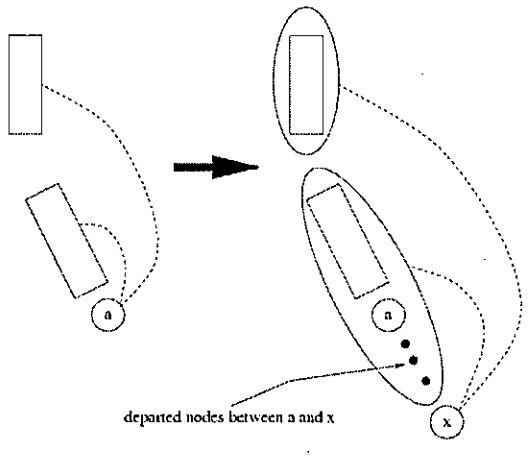


Figure 2. a becomes an internal node.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Case 2: a is an internal node and becomes a junction node because of x.

Suppose that a belongs to a block of internal nodes whose image is z. z may violate Rule 3 for having two junction nodes. If this is the case, it must happen in the upper block of z. Divide the block into two such that each contains a junction node. z takes the block with the nearer junction node, and x takes the other. The non-junction nodes in the original block fall into the two new blocks such that Rule 2 is not violated. Figure 3 shows an example.

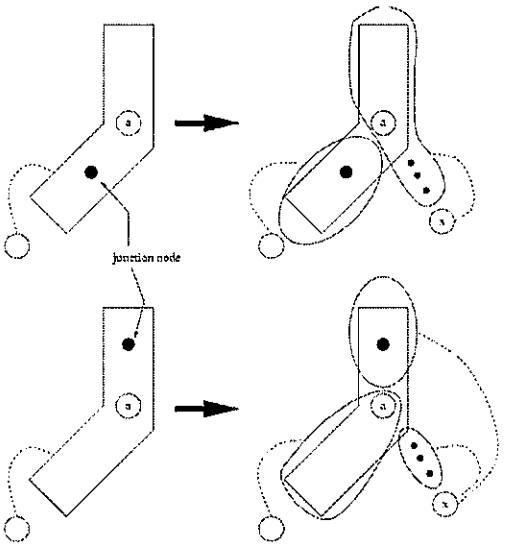


Figure 3. a becomes a junction node.

Maintenance of the junction tree and the subimage tree of a is then needed. A new subimage tree of x is built if a is not the parent of x and there are nodes between them.

Case 3: a is always a junction node.

x will become the image of the departed nodes between a and x, exclusively, as there are no changes to any of the blocks, except that a new block formed by the departed nodes appears with x.

Suppose that a node x leaves the system. The case that x was an internal node before leaving is similar to the one when x joins as an internal node. x's image is informed to take up x's edges. The subimage trees should be maintained.

Consider the case that x is a leaf node before leaving. Suppose that x is not an image node. First maintain the leaf tree, and the free-leaf tree. Second, if x's nearest ancestor has only two child edges, it will become a non-junction node after x departs. Then the junction tree must be maintained. If the ancestor belongs to an upper block of internal nodes, Rule 2(e) will be violated as there will be no junction node in a non-empty upper block. To keep this rule, this upper block will be combined with the block immediately below it, which involves image swapping.

Consider that x is an image node. It is necessary to find another leaf to be the image of the remaining internal nodes. We assume that x has two blocks of internal nodes and at least one of them is non-empty. If the lower block is non-empty and contains one subimage, we choose the last subimage to take all the duties of x. If the lower block contains no subimage, then the whole block will be destroyed as x leaves, because its existence is for x only. Now consider the upper block, if it exists. Divide it into two blocks such that the junction node w belongs to the top one. The bottom one (without w) will combined with the block immediately below it. Image swapping is then needed. The top block will find a free leaf from the descendants of w, which can be done by searching the free leaf. The existence of a free leaf in the subtree rooted at w is guaranteed by the tree structure. After a free leaf is found, pass the edges from x to the new leaf, maintain

ISBN: 978-1-941968-02-4 ©2014 SDIWC

the trees, and it is done. The discussion of the leave-and-join operations is complete.

When a source node wants to send an invitation to a destination node, it first uses its subimage tree. If the destination is in the same subimage tree, it is done. Otherwise, go to the top subimage and reach the first junction node above it. Then use the junction tree to search for the junction node nearest the destination. If the destination is not the junction node, then use a subimage tree to reach the destination, if it exists.

Lastly, we shift back to discuss the search for the nearest online ancestor. When the node x wants to leave, or to simply find its online ancestor, the join algorithm can be used to find the nearest junction node y, and then, one of the online subimages (if it exists) between x and y is x's nearest online ancestor. If there is no such online subimage, the nearest online ancestor is y if y is an online junction node. If y is offline, then search for the one above y. The number of messages used is bounded by O(log N) as there are O(log N) junction nodes, including the offline ones, and at most one subimage tree will be traversed.

7 CONCLUSION

We have given a novel P2P network with less network traffic for future collaborative applications with a static hierarchy of maintained knowledge. One drawback can be that the number of edge changes is $O(\log_k N)$ in the worst case. However, we conjecture that the number of edge changes will be far below this in practice. Our future work is to explore the parallelism of our approach, and verify them by experimental work.

ACKNOWLEDGEMENT

This conference version is supported by Conference Support (UDP) No. 47820 from BAP in Istanbul University.

We thank the reviewers for their valuable comments.

7 REFERENCES

[1] S. Ahmad, C. Bouras, E. Buyukkaya, R. Hamzaoui, A. Papazois, A. Shani, G. Simon, and Z. Fen. "Peer-to-Peer Live Streaming for Massively Multiplayer Online Games," IEEE 12th International Conference on Peer-to-Peer Computing (P2P), September 2012, Tarragona, Spain.

[2] E. Bulut and B. K. Szymanski. "Constructing Limited Scale-Free Topologies over Peer-to-Peer Networks," {\em IEEE Transactions on Parallel and Distributed Systems}, to appear.

[3] I. Clarke, O. Sandberg, B. Wiley and T.W. Hong. "Freenet: A distributed anonymous information storage and retrieval system," Workshop on Design Issues in Anonymity and Unobservability, 311--320, July 2000, ICSI, Berkeley, CA, USA.

[4] L. Fan, P. Trinder, H. Taylor. "Design issues for Peer-to-Peer Massively Multiplayer Online Games," International Journal of Advanced Media and Communication, Vol. 4(2), March 2010.

[5] Gnutella Specification v0.4. "http://www9.limewire.com/developer/gnutella_protocol_0.4.pdf}."

[6] B. Knutsson, H. Lu, W. Xu, and B. Hopkins. "Peer-to-Peer Support for Massively Multiplayer Games," Procedings of INFOCOM 2004, 2004.

[7] X.A. Li and C.G. Plaxton. "On Name Resolution in Peer-to-Peer Networks," The Proceedings of the Workshop of Principles of Mobile Computing, 82--89, October 2002, Toulouse, France.

[8] S. Ratnasamy, P. Francis, M. Handley and R. Karp. "A Scalable Content-Addressable Network," SIGCOMM'01, August, 2001, San Diego, California, USA.

[9] S. Rieche, K. Wehrle, M. Fouquet, H. Niedermayer, L. Petrak, and G. Carle. "Peer-to-Peer-Based Infrastructure Support for Massively Multiplay Online Games," The 4th IEEE Consumer Communications and Networking Conference (CCNC), January, 2007, Las Vegas, Nevada, USA.

[10] A. Rowstron and P. Druschel. "Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems," The 18th IFIP/ACM International Conference on Distributed Systems Platforms (Middleware 2001), November 2001, Heidelberg, Germany.

[11] R. Schollmeier. "A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications." Proceedings of the First International Conference on Peer-to-Peer Computing (P2P'01), Sweden, August 2001.

[12] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek and H. Balakrishnan (reverse alphabetical order). "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," ACM SIGCOMM 2001, San Deigo, CA, August 2001.

[13] J. Walkerdine, D. Hughes, P. Rayson, J. Simms, K. Gilleade, J. Mariani, and I. Sommerville. "A framework for P2P application development," Computer Communications, vol. 31, 387--401, 2008.

[14] L. Xiao, Y. Liu, and L.M. Ni. "Improving Unstructured Peer-to-Peer Systems by Adaptive Connection Extablishment." IEEE Transactions on Computers}, vol. 54, 9, 1091--1103, 2005.

[15] B.Y. Zhao, L. Huang, J. Stribling, S.C. Rhea, A.D. Joseph, and J. Kubiatowicz. "Tapestry: A Resilient Global-scale Overlay for Service Deployment," IEEE Journal on Selected Areas in Communications, January 2004, Vol. 22, No. 1, Pgs. 41-53.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Towards Constructing a Platform that Makes Learning Contents on the Web "Anti-Ubiquitous"

Noriki AMANO
Center for Research in General Education, Saitama University
255 Shimo-Okubo, Sakura-ku, Saitama City, Saitama 338-8570, JAPAN
amnrk@mail.saitama-u.ac.jp

ABSTRACT

In this work, we propose a method of constructing a platform that makes learning contents on the Web "Anti-Ubiquitous", resolve the problem of halting study by "ubiquity" in WBL(Web-based learning) and make WBL more effective and beneficial. "Anti-Ubiquitous" that we have proposed is an opposite concept of "Ubiquitous" and means to add constraints to the location and time in "Ubiquitous". Anti-Ubiquitous learning based on the concept increases the awareness and level of concentration of learners and promotes effective and substantive learning. However, it is not efficient for each teacher to create Anti-Ubiquitous learning contents from scratch. Therefore, we propose a method to make existing learning contents on the Web "Anti-Ubiquitous". Currently, since high-quality learning contents such as OCW and MOOC already exist on the Web, the significance of our approach is great.

KEYWORDS

Web-based learning, Anti-Ubiquitous learning, Platform, learning contents

1 INTRODUCTION

Although WBL[1] is ubiquitous learning that is highly convenient (whenever, wherever, whoever), the ubiquity may lead the halting study. This is because it is easy for the priority on learning to drop when there are not any limitations. In fact, it is easy for "anytime, anywhere" to become "sometime, somewhere", as a result, it may lead to the procrastination of learning activities.

In this work, we aim to establish a method to make learning contents on the Web "Anti-Ubiquitous" in order to resolve the problem of halting study by "ubiquity" in WBL and to make WBL more effective and beneficial. "Anti-Ubiquitous" that we have proposed is an opposite concept of "Ubiquitous" and means to add constraints to the location and time in "Ubiquitous". Anti-Ubiquitous learning[2] based on the concept increases the awareness and level of concentration of learners and promotes effective and substantive learning.

However, it is not efficient for each teacher to create Anti-Ubiquitous learning contents from scratch. Because there are a lot of learning contents with high-quality on the Web, it is desirable to realize Anti-Ubiquitous learning by using such existing contents.

From such a point of view, we propose a method to make learning contents on the Web "Anti-Ubiquitous". Concretely speaking, we aim to construct a platform for Anti-Ubiquitous learning on the Web and to adopt the way of adding constraints to the access time and location when we access the learning contents through the platform. Since high-quality learning contents such as OCW(Open Course Ware)[3] and MOOC(Massive Open Online Course)[4] already exist on the Web, the significance of our method is great.

The remainder of this paper is organized as follows: Section 2 mentions our research background and clarifies the problem, which we try to solve. Anti-Ubiquitous learning that we have proposed is also mentioned in this Section. Section 3 mentions our method to

ISBN: 978-1-941968-02-4

make learning contents on the Web "Anti-Ubiquitous" in detail. Section 4 discusses related work. Section 5 concludes this paper.

2 RESEARCH BACKGROUND

2.1 The Inherent Problem in WBL

Nowadays, there are a lot of high-quality learning contents such as OCW, MOOC, etc. on the Web. WBL has spread rapidly all over the world. In fact, several hundreds of thousand students have been rushed to the free online courses provided by famous MOOC platforms like Coursera[5], edX[6], etc.

Such online courses are high-quality and some state-of-the-art technologies are used in their platforms. For example, they provide a place that has the function of SNS(Social Networking Service) for social learning where students mutually learn from each other. In final exams of the online courses, they try to deal with substitute exam takers and identity thieves by analyzing students' typing, etc.

However, even if such MOOC platforms have some state-of-the-art technologies, they cannot handle the inherent problem in WBL, which is halting study by "ubiquity". Although WBL is ubiquitous learning that is highly convenient (whenever, wherever, whoever), the ubiquity may lead the halting study. This is because it is easy for the priority on learning to drop when there are not any limitations. Since it is easy for "anytime, anywhere" to become "sometime, somewhere", it may lead to the procrastination of learning activities[7].

Actually, from usage situations of the learning management system in our institution[1], we had to understand the reality that many students had accessed to the learning contents only one week just before final exams.

2.2 Anti-Ubiquitous Learning

We have proposed Anti-Ubiquitous learning that is the learning "at specified time and place" by using ICT(Information and Communication Technology) to create such a virtual situation for learners. It is based on the directly opposite concept from "Ubiquitous", however the foundation is "e-learning". Anti-Ubiquitous learning is not "Non-Ubiquitous" one that does not use e-learning at all, but "Anti-Ubiquitous" one that is produced by adding restrictions and limitations to ubiquity in e-learning.

Anti-Ubiquitous learning makes it possible for learners to learn only at specified time and place by themselves. Therefore, since the learners feel that "we can only study now and here", Anti-Ubiquitous learning increases the awareness and level of concentration of learners and promotes effective and substantive learning. As a result, the learners develop self-motivated learning attitude and regular learning habits, because they specify "learning time and place" by themselves.

In Anti-Ubiquitous learning, the most important point is the constraint and/or restriction on the learning time and place. This is because the time and location are thought to occupy an extremely important position during learning. In fact, the times where people are able to concentrate are various greatly between individuals depending on lifestyle habits and natural biorhythms. Based on these points, more effective learning can be anticipated by choosing the time and place of learning carefully.

However, it is difficult to realize Anti-Ubiquitous learning completely because of technical issues such as accuracy of location information, etc. Nevertheless, assuming that there is some degree of correlation to the time and place of learning, it is possible to simulate Anti-Ubiquitous learning by using existing LMS(learning management systems) in a pseudo way. Actually, we practiced pseudo Anti-Ubiquitous learning in real lectures by using WebClass[8] that is an existing LMS, and verified the effectiveness[9]. In particular, we practiced two methods in two classes of the

[1] It indicates Okayama University that is the author's former institution.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

same subject. One is pseudo Anti-Ubiquitous learning, and the other is ubiquitous learning. We compared the average score and time of learning in both classes. On both the score and the time of learning, the class in pseudo anti-ubiquitous learning was superior to those of the class in ubiquitous learning.

Furthermore, we designed and implemented a prototype system for Anti-Ubiquitous learning. The prototype system is based on LMS, which is a platform of e-learning. It enables or disables the accesses of learning contents according to the specified time and place of each individual learner. Although it was built on the Web, we had to implement learning contents according to the specific specification for Anti-Ubiquitous learning from scratch. Moreover, there was a fatal problem, which we could not use a lot of existing useful learning contents on the Web by using the prototype.

3 A METHOD FOR ANTI-UBIQUITOUS

3.1 Basic Policies

We consider the research background in Section 2, and aim to establish a method that makes existing learning contents on the Web "Anti-Ubiquitous". Specifically, towards learning contents on the Web, we design a mechanism for adding constraints about the learning location and time, and implement a prototype system. The basic policies are the followings:

1. It does not alter the implementation (HTML, CSS, etc.) of learning contents.
2. It does not require any special knowledge and skills to users.
3. It does not require special programs except for Web browsers.
4. It does not depend on particular Web browsers.

The above 1 and 2 are prerequisites rather than policies. The above 1 also includes the change of configuration files of Web sites that have learning contents. The above 3 means a special program except for Web browsers, but there may be some cases where plug-ins for Web browsers will be required. Although we aim to realize the above 4 as far as possible, it is difficult at present. Since mechanisms for HTML5[10], CSS3[11], etc. are not implemented by some Web browsers, it is hard for us to achieve the above 4 completely.

3.2 Difficult Problems

The most difficult problems on Anti-Ubiquitous learning contents on the Web are the followings:

- hiding URLs of learning contents
- real-time control of learning location and time

In this work, to make learning contents on the Web "Anti-Ubiquitous" means to add restrictions to accesses of learning contents according to the time and location. However, if we cannot hide URLs of learning contents, the effect of "Anti-Ubiquitous" is reduced by half. Needless to say, if URLs are known, we can access directly to learning content at anytime, anywhere. This is a fatal problem, but from the security point of view, to hide the address bar of Web browsers is inadequate. Currently, it is a specification that we cannot hide the address bar in many Web browsers.

Using the redirection of HTTP, it is possible to realize Anti-Ubiquitous learning contents without hiding URLs[12]. However, in that case, there is a need to modify configuration files in Web sites where there are learning contents. It is a violation of the basic policy 1 in Section 3.1, and it is impossible in the real.

The access control of learning contents is required not only at the start of learning, even during learning. The time is changing, and learners may move during their learning. In other words, from the beginning to the end of learning, we have to continue monitoring the time and location of learners, and if constraints

ISBN: 978-1-941968-02-4 ©2014 SDIWC

on access time and location are no longer met, we must block access to learning contents.

However, the exact track of learning location is difficult with current technology. In this work, we rely on technological innovation in the future about this. The accuracy of location information measurement stays in the range of current technology.

3.3 Handling the Problems

It is difficult to solve the problems in the previous section completely. At present, we are working on the implementation of a prototype system based on the followings.

- Use of inline frame in HTML5
- Prohibition of right-click on Web pages
- Use of Ajax for asynchronous communication

If we put learning contents into inline frame in HTML5, URLs of the learning contents are hidden outwardly. However, if right-click on Web pages is available, the URLs are revealed from source codes by using the context menu on Web browsers. Therefore, right-click on Web pages must be prohibited. This can be easily implemented in JavaScript. Unfortunately, these methods are incomplete at all. This is because it is also possible to display source codes of Web pages from the menu of Web browsers. Even if we can hide the address bar and menu of Web browsers, it is insufficient. We must also prohibit screen capture and printing. Although we can do such things, what can we do for screen photography with digital cameras? For these points, there are researches of prevention technologies of spying displays[13]. In order to exceed the scope of this work, we do not adopt such technologies.

Moreover, we use Ajax to control access for location and time during learning. By using Ajax, changes of learning location and time can be acquired asynchronously from Web browsers. As a result, we can control the access to learning contents on Web sites.

3.4 System Architecture of Prototype

We are currently working on the implementation of a prototype system for Anti-Ubiquitous learning. Figure 1 shows the system architecture.

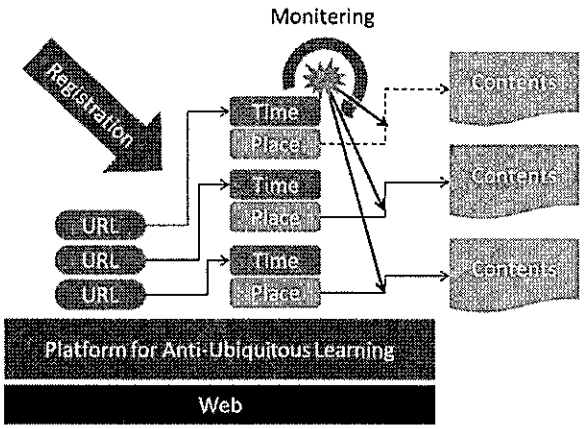


Figure 1. System Architecture of Prototype

The prototype system is a platform on the Web. When users access learning contents through the platform, the accesses are controlled depending on the learning location and time. The followings are procedures of use.

1. User registration (initial time only)
2. Registration of names and URLs of learning contents
3. The learner selects a registered learning content and registers her or his learning location and time.
4. The learner logs in the platform from the registered location at the registered time, then she or he studies by using the learning content.
5. When the learner leaves the specified location during learning or the time of learning is over, the platform will detect it and block access to the learning content.

As the above, the mechanism of the platform is simple and obvious. Currently, we do not handle the access control of learning contents in

real-time for frequent movements of users during learning. Namely, after learners are away from their learning locations and their accesses are blocked, if they return to their locations during learning, they must log in again.

In addition, we let the platform have the function of storing learning histories of registered learners. As we described in Section 3.2, the complete hiding of URLs of learning contents is too difficult. Therefore, the learners can access learning contents directly without using the platform. In such case, there is no meaning to build a platform like this.

However, the purpose of "Anti-Ubiquitous" is not to prohibit learning. It is to boost the awareness of students who cannot learn proactively. Paradoxically speaking, it is not necessary for students who are accustomed to learn proactively to use this platform. But if they do not use the platform, their learning histories do not remain in the platform. This means there is no evidence that they have learned. From teachers' point of view, they can see learning histories of students in the platform and can use the histories for academic assessment. By devising how to use the platform in this manner, we can avoid the fatal problem such as incomplete hiding of URLs.

3.5 The Role and Authority of Users

As users of this platform, we assume teachers and students. Now, we examine whether we distinguish the roles of teachers and students. The issue is the following.

"Should we permit students to register learning contents into the platform?"

In terms of proactive learning support, it is desirable that students can also register learning contents into the platform. However, in such case, to hide URLs of learning contents becomes almost meaningless. As described above, it is not essential to hide URLs of learning contents. However, it is unlikely that students without regular learning habits and proactive learning attitude to register learning contents into this platform and to practice autonomously Anti-Ubiquitous learning.

In addition, the platform needs a mechanism to bind teachers and students. To see and evaluate learning histories of students by teachers compensates the incomplete hiding of URLs of learning contents and promotes the use of the platform by students.

4 RELATED WORK

In e-learning such as WBL, since the ubiquity that gets rid of constraints of learning time and location is the merit of the best, there is no idea such as addition of constraints to the learning time and location. However, as similar methods, there are SRL(Self-Regulated Learning) and CBL(Cohort-Based Learning).

SRL[14] is a learning method that learners specify schedules of learning, study according to the schedules, and evaluate learning outcomes by themselves. As a method to enhance the effectiveness of e-learning, SRL comes to the front recently. Although SRL is similar to Anti-Ubiquitous learning in terms of initiative learning schedule by learners, it is hard to bring a sense of tension for learning by specifying the schedule only. SRL does not have the concept of learning location, either.

CBL[15] is a group learning method that designates start and end points of a learning course, and requires learning targets and problems given at fixed intervals to be cleared. University of Illinois adopts CBL in order to increase the effectiveness of e-learning, and obtains high achievements. CBL shares several things in common with our work in terms of setting temporal constraints on e-learning. However, in Anti-Ubiquitous learning, time settings are based on regular learning habits of learners. Moreover, CBL does not have the concept of learning location, either.

Kajita has been doing the research of a context-aware LMS(Learning Management System)[16]. It is based on ubiquitous

ISBN: 978-1-941968-02-4 ©2014 SDIWC

computing. It provides educational services for students and faculty staffs depending on their context. Therefore, it can be regarded as a kind of research on Ubiquitous Learning[17]. There is also similar work by Li[18], etc. Ubiquitous learning and Anti-Ubiquitous learning are not competitive but complementary.

For the problem of halting study in e-learning, there are also methods which are not systematic: mentors in e-learning[19], Blended Learning[20] that combines face-to-face lectures and e-learning, etc. However, students in MOOC are more than hundreds of thousands, and the number of students in WBL is uncountable. In such case, the introduction of mentors in e-learning and/or the realization of Blended Learning are not appropriate and realistic.

5 CONCLUSION

In this paper, we proposed a method that makes learning contents on the Web "Anti-Ubiquitous". Since a lot of high-quality learning contents such as OCW and MOOC already exist on the Web, the significance of our method that realizes Anti-Ubiquitous learning by using such learning contents is great.

We are in the stage of implementing a prototype system using PHP and HTML5. In the prototype system, we use Geolocation API[21] to acquire location information. Since the accuracy of the location information depends on each Web browser, we are considering to focus on specific Web browsers at the moment.

Furthermore, we have a plan to use the platform in our classes. After carrying out the experiment and evaluation of the platform in real, we open it in public.

Acknowledgment

This work is supported by JSPS KAKENHI (Grant-in-Aid for Scientific Research(C)) Grant Numbers 26330400.

REFERENCES

[1] H. F. O'Neil, R. S. Perez, "Web-Based Learning: Theory, Research, and Practice", Routledge, 2006.

[2] N. Amano, "Anti-Ubiquitous Learning: A New Learning Paradigm", Proc. of the Ninth IASTED International Conference on Web-based Education (WBE 2010), pp.219-224, 2010.

[3] OpenCourseWare, http://www.oeconsortium.org/

[4] M. Nanfito, "MOOCs: Opportunities, Impacts, and Challenges: Massive Open Online Courses in Colleges and Universities", CreateSpace Independent Publishing Platform , 2013.

[5] Coursera, https://www.coursera.org/

[6] edX, https://www.edx.org/

[7] C. Kogo, A. Nakai, E. Nozima, "Relationship between procrastination tendency and student dropouts in e-learning courses", Japan Society for Educational Technology, JSET04-5, pp.39-44, 2004. (in Japanese)

[8] WebClass, http://www.webclass.jp/ (in Japanese)

[9] N. Amano, "An Experiment and Consideration of Pseudo Anti-Ubiquitous Learning by using Learning Management System WebClass", Journal of Japan e-Learning Association, Vol.13, pp.87-94, 2013. (in Japanese)

[10] HTML5, http://www.w3.org/TR/html5/

[11] CSS3, http://www.w3.org/Style/CSS/

[12] D. Gourley, B. Totty, M. Sayer, "HTTP: The Definitive Guide", O'Reilly Media , 2002.

[13] T. Yamada, S. Goshi, I. Echizen, "Method for Preventing Illegal Recording of Displayed Content Based on Differences in Sensory Perception between Humans and Devices", Journal of IPSJ Vol.54, No.9, pp.2177-2187, 2013. (in Japanese)

[14] B. J. Zimmerman, D. H. Schunk, "Self-Regulated Learning and Academic Achievement: Theoretical Perspectives", Lawrence Erlbaum Associates, 2001.

[15] I. M. Saltiel, C. Russo, "Cohort Programming and Learning: Improving Educational Experience for Adult Learners", Kieger Publishing Co, 2001.

[16] S. Kajita, R. Iwasawa, T. Kanegae, S. Ura, A. Nakazawa, K. Kakusho, H. Takemura, M. Minoh, K. Mase, "Development of Context-aware CMS under Ubiquitous Computing Environment", 8th Annual WebCT User Conference, Chicago, IL, 2006.

[17] T. T. Kidd, I. Chen, "Ubiquitous Learning: Strategies for Pedagogy, Course Design, and Technology", Information Age Publishing, 2011.

[18] L. L. Zheng, Y. Ogata, H. Yano, "A conceptual Framework of Computer-Supported Ubiquitous Learning Environment", International Journal of Advanced Technology for Learning, Vol.2, No.4, pp.187-197, 2005.

[19] O. Simpson, "Supporting Students for Success in Online and Distance Education : Third Edition" , Routledge, 2013.

[20] J. Bersin, "The Blended Learning Book: Best Practices, Proven Methodologies, and Lessons Learned", Pfeiffer, 2004.

[21] Geolocation API Specification, http://dev.w3.org /geo/api/spec-source.html

Extracting Agent-based Models for Considering Cultural Factors using Multilingual Case Method System

Kenji Terui and Reiko Hishiyama
Graduate School of Creative Science and Engineering,
Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555 Japan
{terken-06@ruri., reiko@}waseda.jp

ABSTRACT

Globalization has resulted in the need for businesspeople to have a global perspective. Various organizations are actively promoting the development of educational institutions to create diverse classrooms as an effective environment for the helping students to become global businesspeople. Recently, online educational environment has gained significant attention in the information technology development. Educational institutions all over the world have gained increasing research interest in Massive Open Online Courses (MOOCs), and these courses are described as an effective solution for problems associated with student diversity. In this study, we extract agent models based on cultural characteristics to construct an environment that simulates the student diversity. We conducted participatory case study experiments and achieved state transition diagrams that are related to the characteristics of each nationality. Finally, we analyzed their differences and similarities. By comparing the experimental results of culturally different participants, we found that we could extract differences of protocol patterns among the participant groups: American, South Korean, and Japanese. We determined the possibility of creating cultural related player agents by the application of cultural differences that we identified as important for the protocol description required for constructing agent-based models.

KEYWORDS

multilingual case study system; multicultural collaboration; cultural diversity; online educational environment; agent-based models;

1 INTRODUCTION

Owing to the increase in globalization, businesspeople now need to have a global perspective. Various organizations actively promote the education of global human resources and the internationalization in Japan. In addition, international collaboration between universities has resulted in more foreign students studying in Japan. Therefore, in the various educational fields that include case studies, students are expected to gain new knowledge through the synergistic effect achieved by teams comprising members from different cultures or nationalities [1].

However, students who are from different cultures and speak different languages need to overcome non-native language (common language in participants) communication and the lack of cross-cultural understanding to accomplish mutual understanding. Language and culture cannot be represented by the statement "If it differs in the speaker's culture, it differs not only in terms of the content, but also in the manner in which it is said" [2][3]. According to Omi [2], it is necessary to understand the different thinking patterns and the differences in perception. Further, a culture cannot be understood or inherited unless it is experienced. In fact, for global human resource education, we need to understand and experience other cultures through methods such as communication.

Our research goal is to extract agent-based models by considering cultural factors to create an environment that can be used to easily experience cross-cultural aspects. Thus, we extract behavior models through participatory experiments and analyze the characteristics of each nationality. Then, we construct state transition diagrams for agent-based models. An agent-based model is a class of computational models used to simulate the actions and interactions of autonomous agents for assessing their effects on the system as a whole. Agent-based models consist of dynamically interacting rule-based agents. The systems within which they interact can create real-world-like complexity. State transition diagrams allow the easy construction of classroom diversity based on the extracted agent-based models. In addition, agents can autonomously determine their behavior while interacting with the class environment and other agents. Multiagent diverse classrooms represent individual decision-making in detail according to each agent's circumstances and reproduce complex phenomena that arise from case discussions between different agents, which is an educational benefit for

ISBN: 978-1-941968-02-4

the participants[4].

2 PROPOSED APPROACH

The case method is a teaching approach that consists of presenting the students with a case and placing them in the role of a decision maker facing a problem [5]. Presently, students that access the Internet participate in online courses, and take classes and discuss case studies with other students. However, case materials used in MBA programs are mainly written in English. Therefore, the students are required to be to understand and speak fluent English along with their native languages. Therefore, it is difficult for students who are non-native English speakers to read and participate in case discussions. In general, non-native English speakers need to improve their English skills to participate in MBA programs for a few years of study abroad. The language classes that need to be attended out of necessity are expensive for the students. Therefore, to avoid this additional expenditure, it is necessary to resolve this issue.

In this study, we prepared an online class environment in which students can participate in the case method using their own native language. In this system, students from many different countries speak with other participants and discuss case materials with respect to different cultural backgrounds. This class can overcome the linguistic barrier at some level and enable students to easily participate in cross-cultural discussions. The left-hand side of Fig. 1 describes a recent online class style in which students from different countries attend the class at a particular time and participate in case discussions. However, it is practically difficult for students to attend a class at a particular time owing to physical barriers such as time differences. Therefore, the case method in a diverse environment is hampered by physical barriers. Our proposed approach can easily maintain a diverse environment by overcoming the abovementioned barriers (Fig. 1). The right-hand side of Fig. 1 describes replacement participants as agents. We adopt an agent that considers cultural factors instead of participants that have different cultural backgrounds. Agents that consider each cultural characteristic such as thinking and behavioral mannerisms participate in the discussions. The system provides an educational environment that students can use without the influence of time differences. However, to design these agents, we need to design a large cultural variety of agents. We can create an environment consisting of a variety of (virtual) students by preparing diverse agents by simulating the case study class in the real world. In particular, in the case of dealing with cultural problems in business, this system produces the effect expected in the above environment with a high accuracy. Therefore, we collect the student's behavior model through actual discussions about business case materials. Then, we extract the agent model to create a student agent that considers cultural characteristics for each nationality.

2.1 Extraction Procedure

We extract agent models as per the method stated in [4]. To obtain cultural characteristics related to a agent model, we conduct a simulated case method for use in an actual education field, and collect the discussion logs.

We apply the results of the analysis of discussion logs to protocol analysis to reveal the process of communication for all participating nationalities. First, we conduct a simulated case study using participants from different nationalities using the multilingual case system, which is explained in Section 3.1. Next, we collect discussion logs that is derived from each nationality's participants using real case discussions. Finally, we assign utterance tags to the discussion logs for extracting agent models (Fig.2).

The tags used in Fig. 2 are summarized in Table 1. We divide the discussion logs into short topics and assign utterance tags on the basis of utterance types (Table 1) to each utterance using the method proposed by Wang et al. [6] for discussion logs.

Table 1: Utterance types (Tags)

Type	Definition
Opinion	Own opinion about case material
Agree/Disagree	Agreement/disagreement with other participants
Question	Utterance asking for feedback
Explanation	Supplemental explanation about own opinion
Procedure	Mainly facilitator's utterance
Other	Unclassifiable utterance

We observe the condition of each participant using tag information and extract some states. Similarly, we extract utterance types that change from their current state to the next state. Finally, we tally the data described above.

Further, we construct a state transition diagram using the method proposed by Torii et al. [7] in which the discussion logs are divided into minimum discussion topics. We also analyze the behavioral tendencies of each participating nationalities. Finally, we extract agent-based models considering the cultural factors through the application of the above data counted for each nationally based on protocol description.

3 EXPERIMENT

We conducted a participatory experiment in which we applied our simulated case method [8]. We conducted the experiment in accordance with the process of the case method. First, each participant logged in, selected their native language, and read the case mate-

ISBN: 978-1-941968-02-4 ©2014 SDIWC

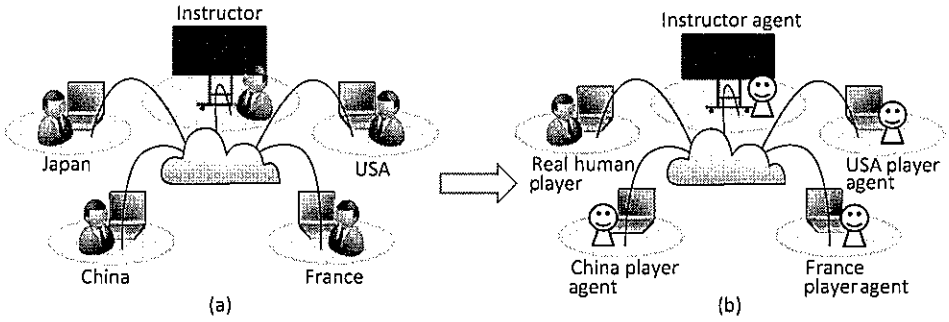


Figure 1: Class images: (a) recent style and (b) proposed style

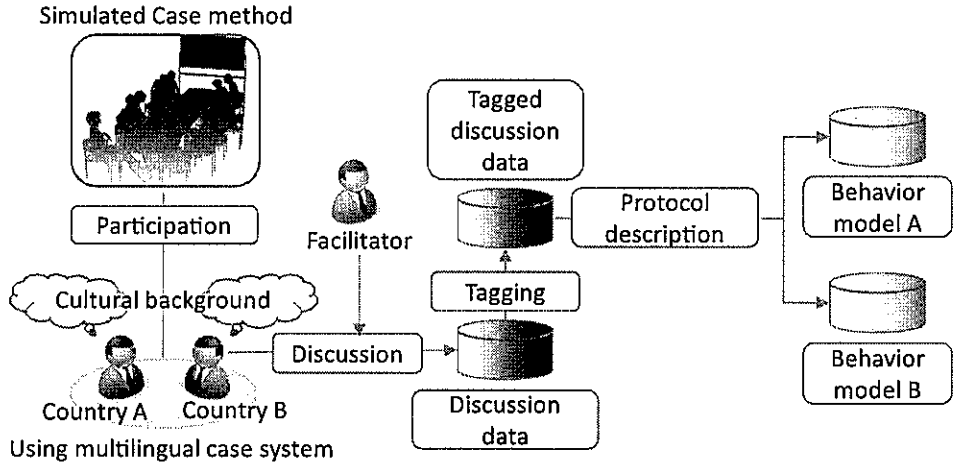


Figure 2: Procedure for the extraction of an agent model

rials shown on the Web. The selected case was described in their native language. Next, they chose a sentence in the case content and noted it with a line number and case passage using analysis notes for organizing information. The participants then estimated the meaning of the sentence by back-translation, although they worked in their own language. Participants' analysis notes are stored in the shared analysis note database. Then, the participants discussed the problems in the case using multilingual chat. Text messages were stored in the Chat database. After the case discussion, we conducted a questionnaire regarding the participant's opinion about each nationality, mistranslations, this system, and the case method. We conducted cross-cultural analysis using the discussion logs. We applied the discussion logs recorded dynamically during the process of interactive communication between the participants to analyze the thinking processes and behavior.

3.1 Experiment System

Figure 3 shows the configuration of the system used in the experiment. This system [9] was developed on a cloud-based computing environment and was connected to the Language Grid [10]. We used the "multilingual studio" [11], which is a set of application

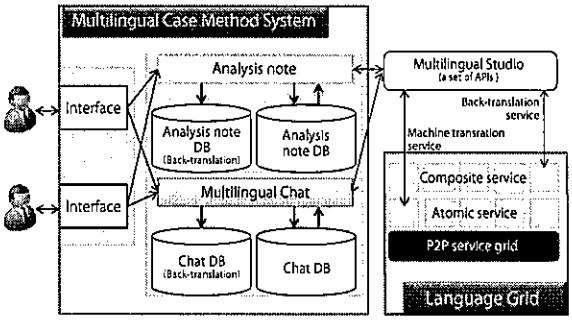
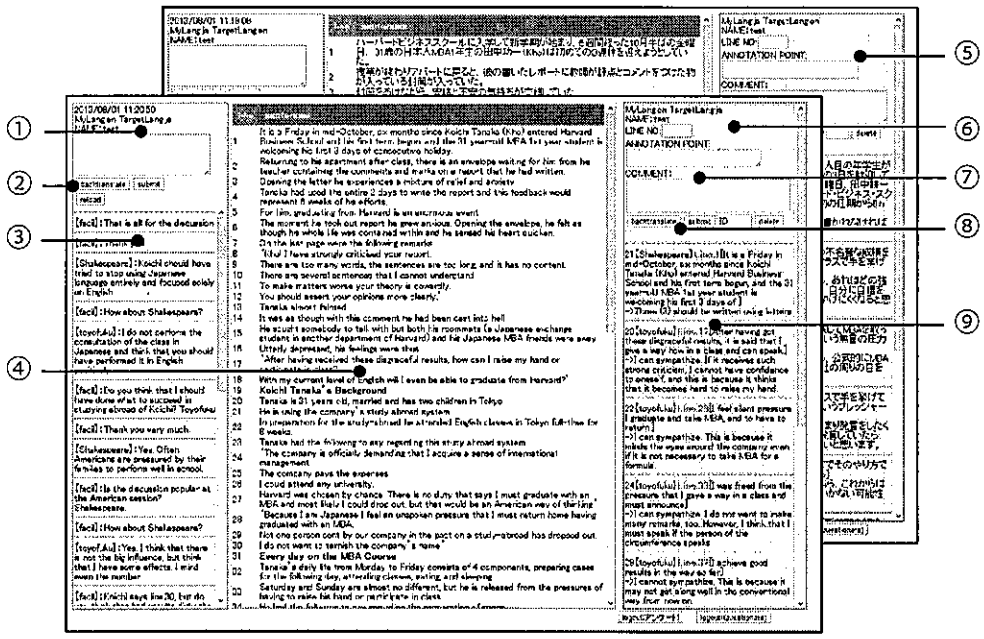


Figure 3: System configuration

programming interfaces (APIs) for using the multilanguage Web services, which are language sources (e.g., machine translation and dictionaries), provided by the Language Grid. We can shift between more than 170 language services (e.g., French, Vietnamese, and so on) by changing the language service API based on the case language.

3.2 Participants

Eight pairs of university students belonging to the United States, South Korea, and Japan participated in the experiment (six Japanese speakers from Japan, six



①Multilingual chat function
②Back-translation button
③discussion logs
④Case contents
⑤Interface for Japanese
⑥Analysis note function
⑦Participants Write analysis note with line No.
⑧Back-rtanslation button
⑨Analysis notes of participants

Figure 4: Application interface (The bottom-left is an interface for American English and the top-right for Japanese)

South Korean speakers from South Korean, and four English speakers from the United States). We conducted experiments using three pairs of two Japanese students (six Japanese students), three pairs of two South Korean students (six South Korean students) and two pairs of two American students (four American students). Meanwhile, a Japanese national acted as a facilitator in the experimentation. This experiment plan was based on Yamashita's experiment plan [12]. Her experiments were conducted using six pair of students, and she analyzed the conversation data using the conversation log. Using the same conditions as her experiments, we conducted the experiments using Machine Translation(MT)-mediated case discussion experiments.

4 RESULT

The number of discussion logs obtained through experiment is as follows: American students have 63 discussion logs, Japanese have 136 discussion logs, and South Korean have 122 discussion logs. We show the example of tagging discussion logs and state transition of participants in Fig. 5.

The obtained state transition process is as follows: Participants transit to state A(Information arrangement) in which a participant gives his/her own opinion at the facilitators direction. If a participant cannot understand the meaning of a facilitators direction or the other participants utterance (because of mistranslation), the participant requests an explanation and transits to state C(Waiting for explanation). When a participant wishes to know the other participants opinion, he/she transits to state B(Waiting for other opinion by asking a question).

We recognized four patterns of opinion-making behavior.

1. opinions stated after a facilitators direction
2. opinions that agree/disagree with other stated opinions
3. opinions that answer a question
4. opinions that explain or additionally supplement one's own opinion

Behavior patterns 1, 2, and 3 reveal one's own opinion. On the other hand, we also find instances of pattern 4.

Moreover, the flow of discussion begins to disrupt after references are passed to each other. Then, the participants scent discussion get stacked up and transit to state D (Waiting for facilitators direction). If a facilitator finds a point of discussion that needs to be explained and informs the participants to do so, they transit to Information arrangement again and discuss the point. If not, the topic is closed and the facilitator introduces a new topic.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Name	Utterance	Tag	State transition	
			Now	Next
facili	Then, do you think that there is a relationship to speak of in class and report to Lowest Score?	Procedure	Start	
1	I get the impression that he was the only foreign student in the class, so understandably his score would be the lowest.	Opinion	A	A
2	No because there may be some people who are shy in class and understand what the teacher is saying. If not student isnt willing to speak up because the person doesnt understand the material, then yes there is a relationship between the two.	Opinion	A	A
1	But speaking up in class does help you tell if you are on the right track and thus helps you produce better results on any assignment.	Agree/Disagree	A	A
2	Thats true. But if the person is too shy to speak in class, the person could always go to the teachers office hours or research it himself.	Agree/Disagree	A	A
2	Or.. What Im saying is there are shy people in class who dont speak up but can get good grades.	Explanation	A	A
1	Thats true. I think, in other words, its not about how much you speak up. Its about comprehension. Koichis comprehension was low. That is all.	Agree/Disagree	A	D
2	Yeah. I agree.	Agree/Disagree	A	D
facili	Do you think results and remarks in the classroom that are	Question	turn to A	
1	Not that much. I think its more of a relation between what he wrote on the paper and the results.	Opinion	A	A
1	After all, he only spent 2 days on it.	Explanation	A	D
2	In the case of Koichi, if he did speak up in the classroom, it may not have been beneficial to him. He didnt have a thorough grasp of English to begin with, so his lack of the language was what led him to his bad results.	Opinion	A	D

Figure 5: Example of the discussion log, tagging, and state transition

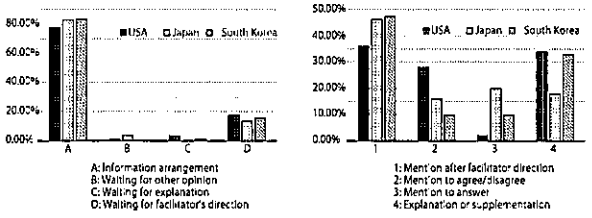


Figure 6: Ratio of state (left) and mention type (right) for every nationality

The counted number of appearance of each state for each participant along with each utterance type is listed in Table 2.

The left-hand side of Fig. 6 shows the ratio of states (A-D in Fig. 7), the right-hand side of Fig. 6 shows the ratio of transition patterns from Information arrangement (1-4 in Fig. 7). In the state graph, state A is the highest of all states for every nationality, followed by D, B, and C. Thus, we found no noted difference. Therefore, conceivably, no cultural difference exists in the ratio of these states. On the other hand, we found a significant difference (chi-square test; $p_i.01$) in ratio of the utterance types (left-hand side of Fig. 6). Utterance types 1 and 4 are active utterances [6] that are opinions that wrap around oneself. Utterance type 2 and 3 are reactive utterances [6] that represent opinions for other's opinions or ideas.

Figure 7 shows a state transition diagram from discussion logs obtained in the experiment. In the figure, the top-left denotes the original transition diagram; the top-right, Japanese; the bottom-left, American; and bottom-right, South Korean. The utterance types from state A in each nationality's state transition diagram reflects cultural characteristics. After counting the ratio of appearances, these lines imply the following: x<15%: thin line, 15% <= x<30%: medium line and 30% <= x: heavy line.

In the case that utterances 2 and 3 have a large appearance, it is possible to regard the nationality as interactive. Americans have a low number of utterance 1; however, the number of utterance 2 is the largest compared to other nationalities. On the other hand, utterance 3 is extremely low for Americans. For the Japanese, we identified that the number of utterance 3 are the largest compared to others. This implies that the Japanese participants constantly explored the same opinion interactively. It would appear that the Japanese tend to say if they agree or disagree after listening to other participants. For the South Korean participants, utterances 1 and 4 account for 80% of all discussions. Thus, South Korean participants are thought to have a less tendency to interact while in a discussion.

5 DISCUSSION

We noted some differences and similarities in the results of the protocol analysis. The Japanese had a high propensity for listening carefully to other participants' points of view. Moreover, the Japanese participants responded to others' opinions and respected their point of view. On the other hand, South Korean partici-

Table 2: Number of state and mention type for every participant

Participant	A	B	C	D	Total		1	2	3	4	Total
American 1	10	0	0	4	14		3	1	1	1	6
American 2	12	0	0	4	16		3	1	0	4	8
American 3	23	1	1	4	29		4	6	0	7	17
American 4	25	0	2	4	31		8	6	0	5	19
Total	70	1	3	16	90		18	14	1	17	50
Japanese 1	30	2	0	2	34		11	4	4	6	25
Japanese 2	28	3	0	4	35		9	4	5	2	20
Japanese 3	27	0	0	5	32		7	4	7	4	22
Japanese 4	29	1	0	5	35		13	2	4	4	23
Japanese 5	10	0	0	3	13		4	2	0	1	7
Japanese 6	12	0	0	3	15		5	1	1	2	9
Total	136	6	0	22	164		49	17	21	19	106
South Korean 1	33	0	0	5	38		13	1	3	11	28
South Korean 2	29	0	0	5	24		11	3	4	6	24
South Korean 3	12	0	0	3	15		5	1	0	3	9
South Korean 4	14	0	0	3	17		8	1	1	1	11
South Korean 5	24	0	0	5	29		9	1	2	7	19
South Korean 6	28	0	2	5	35		7	4	1	9	21
Total	140	0	2	26	168		53	11	11	37	112

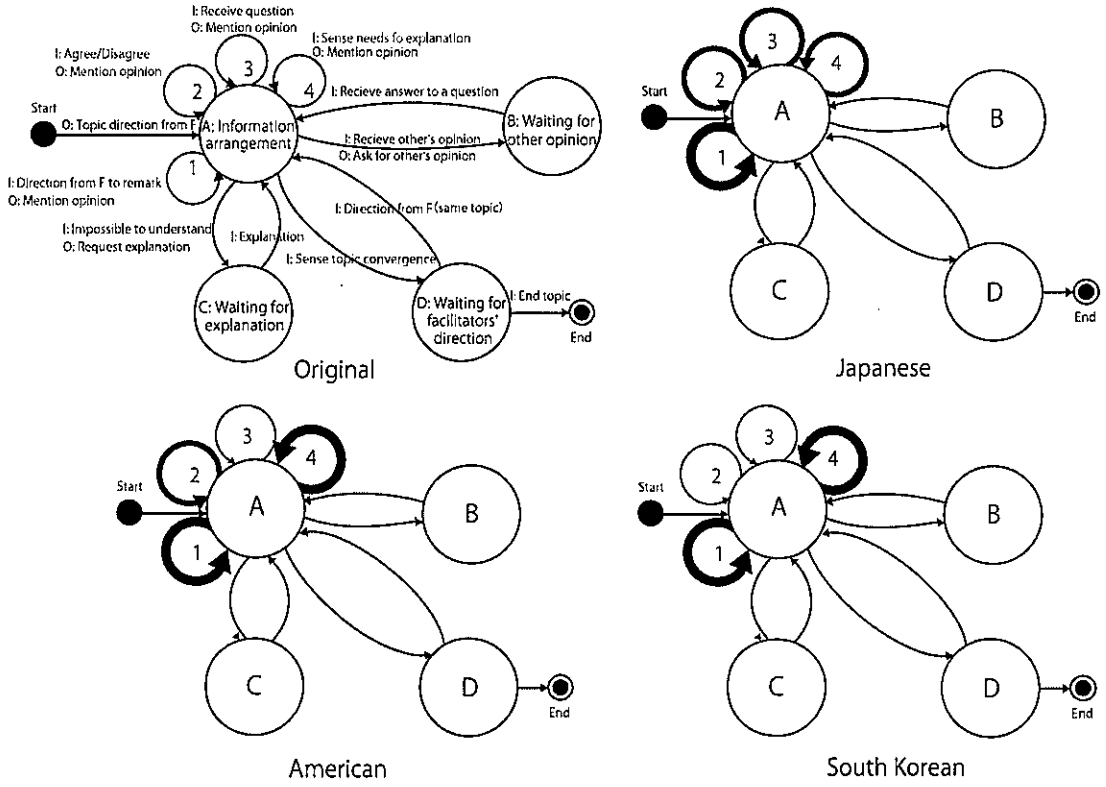


Figure 7: Transition diagram: original, Japanese, American, and South Korean

pants tended to push their opinions more strongly to convince other participants. Further, the Americans tended to argue strongly for their own opinions. However, although they considered other opinions, they tended to disagree with them.

In MT-mediated communication, it is necessary to consider mistranslations. We found a discussion on mistranslation in state C(Waiting for explanation) in the experiment with Japanese and South Korean participants. A facilitator noticed a particular miscommu-

nication between a South Korean and Japanese pair. This facilitator asked the Japanese participant to write in a more understandable manner for the South Korean participant. The South Korean participant understood the message and answered as per the requirement of the Japanese participant. Thus, this helped in overcoming the mistranslation. Therefore, we identified the possibility that third-party support can overcome mistranslations.

We conducted a questionnaire about the amount of mistranslation (scale of one to ten) and communication (five levels: Communicated very smoothly, Communicated satisfactorily, Communicated somehow, Communicated with difficulty, Could not communicate.). According to results of the questionnaire, most participants felt that mistranslation accounted for 21% to 30% of the messages. However, most participants answered "communicated satisfactorily" and "communicated somewhat." Moreover, the "Could not communicate" option was not selected by any participant. In this study, we could not identify the dissonance caused by mistranslation.

6 RELATED WORKS

Since global businesses and organizations increasingly bring people together from around the world to solve common problems [13][14][15], challenges due to linguistic differences occur. Similarly, in the educational context, problems arise from differences in communication styles, relationship norms, negotiation strategies, and methods of dealing with conflict [16]. To solve these problems, the case method was conducted in various educational institutions. The case method was developed in the early 20th century by the Harvard Business School in the United States [8]. Case materials used in MBA programs are mainly written in English, and therefore, it is difficult for students who are non-native English speakers to read and participate in case discussions. Therefore, participants are limited to those who can speak fluent English. There are many business students who attend case discussions using OpenCourseWare even if they cannot speak English.

Recently, the Massive Open Online Course (MOOC)[17] is receiving significant attention. A MOOC is an online course with the option of free and open registration, a publicity-shared curriculum, and open-ended outcomes. A MOOC facilitates the participation of an acknowledged expert in a field of study, and a collection of freely accessible online resources. In addition, it builds on the active engagement of several hundred to several thousand "students" who self-organize their participation according to their learning goals, prior knowledge and skills, and common interests. In general, an MOOC has no fees, no prerequisites other than Internet access and interest, no predefined expectations for participation, and no formal accreditation.

This paper aims to extract agent-based models for students with diverse values and cultural differences. The agent-based models will be available for the case discussion class on the MOOC.

7 CONCLUSION

In this study, we proposed a procedure of extracting agent-based models considering cultural factors, and we extracted agent-based models from actual communication logs by following the proposed process. We conducted a small experiment for three nationalities in this study. On conducting participatory experimentation, we found that the utterance behavior model in a chat discussion can be distilled into a state transition diagram. However, we believe that it is possible to create a player agent that behaves more specific to a particular country. Therefore, we need to extract new characteristics and differences more finely by increasing the number of participants and nationalities.

As our future work, we plan to implement the extracted agent-based models as agents, and conduct experiment by the application of conversation systems using the text corpus.

ACKNOWLEDGMENT

This work was supported by the Service Science, Solutions, and Foundation Integrated Research Program of the Research Institute of Science and Technology for Society, Japan Science and Technology Agency (JST/RISTEX), and by a Grant-in-Aid for Scientific Research (S) (24220002, 2012-2016) from the Japan Society for the Promotion of Science (JSPS).

REFERENCES

[1] J. Miller, A. Kostogriz, and M. Gearon(Eds.), "Culturally and Linguistically Diverse Classrooms :New Dilemmas for Teachers," New Perspectives on Language and Education, 2009.

[2] J. Omi, "Japanese post-secondary ESL students' perspectives on communicative competence in the cultural contexts of the United States," University of San Francisco, 2003.

[3] E.N. Richard, "The Geography of Thought: How Asians and Westerners Think Differently...and Why," Free Press, 2004.

[4] T. Ishida, H. Hattori, and Y. Nakajima, "Multiagent Simulation. Kyoto University Field Informatics Research Group," Field Informatics. Springer, Ishida, T(Eds.), 2012, pp.89-105.

[5] JEREMYCWILSON.COM: Summer Pre-Law Program and My Intro to the Case Method, http://www.jeremycwilson.com/2009/08/summer-prelawprogramandmyintroto_24, last accessed: 2014/7/18.

[6] H-C. Wang, S.R. Fussel, and L.D. Setlock, "Cultural difference and adaptation of communication style in computer-mediated group brainstorming," Proc. CHI 2009, NY: ACM Press, 2009, pp.667-678.

[7] D. Torii, T. Ishida, S. Bonneaud, and A. Drogoul, "Layering social interaction scenarios on environmental simulation," In Multi-Agent and Multi-Agent-Based Simulation, Springer Berlin Heidelberg, 2005, pp.78-88.

[8] E.M. Morgan, "The Case Method," Journal of Legal Education, 1952, pp.379-391.

[9] K. Terui, and R. Hishiyama, "Multilingual Case Method System for Cross-Cultural Analysis," Proc. ICCC'13, 2013, pp.117-122.

[10] T. Ishida, "Language Grid, An Infrastructure for Intercultural collaboration," IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06), keynote address, 2006, pp.96-100.

[11] Multilingual Studio Web site: http://langrid.org/developer/jp/index.html, last accessed: 2014/5/14.

[12] N. Yamashita, and T. Ishida, "Effects of machine translation on collaborative work," Proc. CSCW 2006, 2006, pp.515-523.

[13] G. DeSanctis, and P, Monge, "Communication processes for virtual organization," Journal of Computer Mediated Communication, Vol.3, No.4, 1998.

[14] G. Leshed, D. Cosley, J.T. Hancock, and G. Gay, "Visualizing language use in term conversations: designing through theory," experiments, and iterations. Proc. CHI 2010, 2010, pp.4567-4582.

[15] P. Shachaf, "Cultural diversity and information and communication technology impacts on global virtual teams: An exploratory study," Information Management, Vol.45, 2008, pp.131-142.

[16] C.W. Yuan, L.D. Setlock, D. Cosley, and S. Fussell, "Understanding Informal Communication in Multilingual Contexts," Proc. CSCW 2013, 2013, pp.909-921.

[17] A. McAuley, B. Stewart, G. Siemens, and D. Cormier, "The MOOC model for digital practice," 2010, pp.1-64.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Presenting New Method to Optimize Query in Distributed Database System

Sajjad Baghernezhad
Department of Computer, Darab Branch, Islamic Azad University, Darab, Iran
sbaghernezhad@yahoo.com

ABSTRACT

Query optimization is one of the essential problems in centralized and distributed database. The data allocation to different sites is proposed in a distributed DMS(Database Management System) before a query in order to decrease, the next communicative costs namely an optimized bed production which is of 'NP' issues. In this article, it was attempted to examine both the methods to allocate data and produce optimized design in a distributed system and the space to query for query optimization in the distributed environment and show the need concerning optimization method in view of different aspects of optimization process. We install a new method for optimization in distributed database environment which indicates somehow our simple optimization design is executed relatively well until the database design is physical

KEYWORDS

Query Optimization, Distributed Database, Allocation, SQL

1 INTRODUCTION

In recent years the distributed database are used increasingly due to development concerning computer networks and database technology[1]. Distributed database system is dispersed physically but is centralized logically and is a composite of computer networks and database system; generally distributed database technology is the center of different researches such as general integration design, data exchange and query processing and optimization; the query optimization returns to the early distributed systems and recently many researches executed in relation to different potentials of data sources combination and costs model[2]. However, query optimization presents features in a developed distributed environment which changes considerably trade offs in the optimization process. The distributed query processors should take into consideration three essential principles to process and optimize users' query:

Necessary query processing: Processing query is the process to translate a query in a high level language such as SQL to a lower level language. It is possible to access the data and calculate in different sites in a distributed system in big scale so one of the essential goals of the distributed database systems development is to take into consideration some part of a query design in a distributed method to increase efficiency[3].

Necessary costs factors: Considering the tables are in the same place of a centralized database management system the query costs are measured on the base of a one-dimensional factor but in a distributed database it is necessary to control the costs logically by a database by dividing them in different dimensions; usually the response time and computation precision and accuracy are main factors to compute the query costs.

Necessary costs estimation: Considering the tables are among different sites in the distributed systems one of the main goals of distributed systems is to estimate the communicative cost among the sites; of course,

ISBN: 978-1-941968-02-4 ©2014 SDIWC

a centralized optimizer may not estimate accurately the operations' costs in many independent sites. In first section of this article we describe the query optimization and examine essential steps to optimize query in distributed databanks. In the second one we describe optimization architecture and define the problem and in the third one we describe how allocate the data to different sites and methods installing to produce the optimize design and in the following section we examine the proposed method to optimize query and finally we propose essential points for future researches.

2 DEFINING ARCHITECTURE AND RELATED PROBLEM

The most related sections of the system to optimize query are the proposers in independent sites and query optimizer among the devices (Figure). As the query optimizer may use a variety of optimization, algorithm in a centralized database system it is necessary to estimate the cost by essential sources or the fabricated proposers in a heterogeneous database system. The optimizer and relation proposers use two mechanisms:
1 – RFP (Request for proposition) where the optimizer uses one operation.
2 – Proposition by the proposers who estimate the cost (Figure 1).

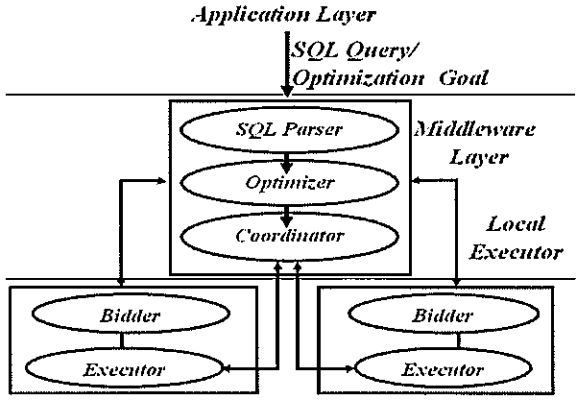


Figure 1. System architecture

3 THE DISTRIBUTED QUERY OPTIMIZATION PROBLEM

The most important issue is related to communicative cost among different sites in the distributed databanks. Contrary to centralized databanks the most important cost is related to time and the memory necessary to execute a query. The distributed query optimization problem is to find an execution design special to the user's query to achieve the goal proposed by the user; such goal may be a function formed by many variables such as response time, total execution cost and data accuracy; for simplicity we focus on two of them: execution time and total execution cost. The optimization issue may be proposed in two general phases in distributed databanks and each one may be examined separately; this section includes the data allocation in different sites to decrease communicative costs or minimize the exchanges in a distributed system and second section is related to produce executive optimized design for a query[4]. A query with link tendency formed of some links may be executed in different designs which has different executive cost but the same result; in continuation we try to describe in detail the two sections

3.1 Data allocation

The most important cost in the executive requests in a distributed database system is the cost for the cost of the data transfer achieved by a request from different site to a site where the request is executed. The data allocation is to define a measure from the frameworks in other sites to decrease total costs appeared during the execution of a collection of the requests. Having executed this method the time mean to execute the request which is very important in a usual distributed group and multimedia database system decreases; however, the problem related to NP data allocation method continues. The execution cost related to the

ISBN: 978-1-941968-02-4 ©2014 SDIWC

request depends on the request and data situation. In view of specialty the data situation in a request defines the amount of the data transfer in processing the request so when someone encounters the data allocation it had better he (she) improves the data. Having defined the collection of the requests achieved from the pieces of the data the pieces are allocated to the sites from the database to decrease the total cost of the data transfer for processing requests. The methods proposed in the field are described. It should be noted that a general form for databank, the method to allocate data and some limits are taken into consideration in some algorithms and some algorithms are stated, but we consider their general state in which the tables of a bank may be divided between several sites. Perhaps there are some copies of a table in several sites or each site may include only one table; with such presupposition we deal with proposed algorithms.

3.1.1 Genetic algorithm

This method select a primary group from division possibilities and enters into genetic cycle as chromosomes shown in an array frame as following algorithm[5]:

```
(1) Initialize population. Each
individual of the
population is a concatenation of the
binary
representations of the initial random
allocation of
each data fragment.
(2) Evaluate population.
(3) no of generation = 0
(4) WHILE no of generation < MAX
GENERATION DO
(5) Select individuals for next
population.
(6) Perform crossover and mutation for
the selected
individuals.
(7) Evaluate population.
(8) no of generation ++;
(9) ENDWHILE
(10) Determine final allocation by
selecting the fittest
individual. If the final allocation is
not feasible, then
consider each over-allocated site to
migrate the data
fragments to other sites so that the
increase in cost is
the minimum.
```

Figure 2. Genetic algorithm

3.1.2 Algorithm to query randomly beside

Main principle in a side searching method is to create a primary solution with medial quality; then based on neighbor defined before it selects a rapid solution in the searching space and tests if it is a better solution or not. If the new solution is better, it accepts its method and begin to query in new neighbor space; otherwise, it selects another solution. The method stop querying after some social steps or the solution stops after passing some stable steps. The quality of querying solution in neighbor space depends on creating neighbor solution; this method is defined to allocate the data as follows:

```
(1) Use Divisive-Clustering [19] to
find an initial
allocation Initial Alloc;
(2) Best Alloc = Initial Alloc;
(3) New Alloc = Best Alloc; iteration
= 0;
REPEAT
(4) searchstep = 0; counter = 0;
REPEAT
(5) Randomly select two sites from New
Alloc;
(6) Randomly select two data fragments
from each
site;
(7) Exchange the two data fragments;
(8) IF cost is reduced THEN
adopt the exchange and set counter to
0;
ELSE otherwise undo it and increment
counter;
UNTIL ++searchstep > MAXSTEP OR
counter >
MARGIN;
```

```
(9) IF cost(New Alloc) < cost(Best
Alloc) THEN
Best Alloc = New Alloc;
(10) Randomly exchange two data
fragments from
two randomly selected distinct
sites from New Alloc; /* Probabilistic
jump */
UNTIL iteration > MAXITERATION;
```

Figure 3. Algorithm querying randomly in neighborhood

3.2 Producing design to execute optimally and related works

In this section we examine the methods producing optimal design to execute a query. The methods producing an optimal design are sorted in two groups based on cost and rule; in the method based on rule essentially the findings are considered from a design in the best link graph and there is no space for a vaster space so this method is rapid and mostly there is no other better one and the algorithms find a better design after one execution. In the method based on cost the base is to apply statistic relation in the estimations and costs; it has query space and uses competency methods for query and finds the best design for little relations and there is no possibility to find design for many relations

3.2.1 Link graph

The query optimization methods are based on that if each database may be considered as a link graph in a way that each node shows a table and each edge shows the relation between the tables, both groups of the algorithms operate by virtue of the link graph(Figure 2)[6].

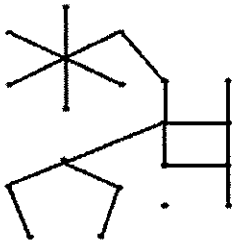


Figure 4. Joint graph.

Algorithms based on rule are the methods with low flexibility and often low efficiency and mostly no optimal design is achieved, but considering they select a design as an optimal design in one execution they are rapid. For instance, we may mention Prim and Kruskal algorithms; these methods are executable only for little databases with limited capacity and practically are not very efficient. In the algorithms based on cost the statistical relations and data in the system catalog are used to estimate costs, etc. And the query space is exponential and competency methods are used for query. The algorithms based on cost are sorted in two definitive and non-definitive groups. In the former algorithm is only a comprehensive and dynamic one and may not reply the great questions, but the non-definitive ones query for a graph whose nodes are alternative executive designs and may be used to reply the question; each node has a cost and the algorithm is to find a node with its least costs[7].

3.2.2 Dynamic programming algorithm

The advantage of this algorithm is to create the best possible design, but its time complexity is multi-phrases and its space complexity is not appropriate to complicated queries; specially in a distributive system the dynamic programming complexity is expensive for many queries[8]; one of the developed states of the dynamic programming algorithm is repetitive dynamic programming creating designs as good as dynamic programming algorithm for and complicated simple queries not available in the dynamic programming. Dynamic programming algorithm is shown for query optimization in algorithm No. 3 operating from down to up and creates more complicated design substructures by simpler ones.

```
INPUTS rels "List of relations to be
joined"
```

ISBN: 978-1-941968-02-4 ©2014 SDIWC

```
OUTPUT pt "Processing Tree"
partialsolutions := {All scans for all
attributes
involved}"Remove all elements from
partialsolutions
with equivalent, lower-cost
alternative"
FOR i := 2 .. |rels|
FOR all pt in partialsolutions
FOR all R in rels such that R not in
pt := pt ∞ R
END
END
"Remove all elements from
partialsolutions with
equivalent, lower-cost alternative"
END
RETURN "Arbitrary element from
partialsolutions"
```

Figure 5. Algorithm dynamic programming

3.2.3 Composite evolution optimization algorithm

Composite evolution optimization algorithm is created and used by composing genetic algorithm, learner's automata ,composing gene and chromosome concepts. The important property of composite evolution algorithm is its resistance against replies' superficial changes. Auto-restoration, reproduction, fine and reward are the composite algorithm features. Contrary to classic genetic algorithms the binary coding or natural overlay exposition are not used in the composite evolution algorithm. Composite evolution algorithm has higher efficiency than the genetic one.

```
Function query optimization (query)
Create the initial population CM1…
CMn;
EvalFitness();
While (Not (Stop Condition)) do
NewCM1 = CM with minimum Value of
Cost;
For i = 1 to n do
Select CM1; Select CM2;
If (Random > PC) then
Crossover (CM1, CM2);
End If
If (Random > PM) then Mutation (CM1);
Mutation (CM2);
End If
NewCMi+1 = CM1;
NewCMi+2 = CM2;
i=i+2;
End For
For i = 0 to n do
CMi = NewCMi;
For i=1 to 4
u = Random *n;
If (costu(CM.LAi)<MeanCost) then
Reward(CM.LAi , u )
Else Penalize(CM.LAi , u );
End If
End For
End For
EvalFitness();
```

Figure 6. Compositive evolution algorithm

4 HOW TO OPTIMIZE THE PROPOSED QUERY

The suppositions to discuss about algorithm and optimization technique proposed in this article are as follows:

Precise statistics: We suppose that there are precise statistics about cardinality and choice; such data may are collected from the standard protocols with permission to query from host database.

Relation costs: We suppose the relation costs are almost stable during optimization and query execution and the optimizer may meet the sustained relation costs in data transfer between two related sites.

There is no tube line throughout sites: We suppose that there is no tube line among the query operators throughout the sites; generally we divide all optimization algorithms in three steps:

1 – Selecting designs' substructures meeting the cost and preparation of the requests for proposals.

2 – Sending the message for the proposers of the request cost.

3 – Estimating costs for the designs and designs' substructures; if possible, to decide how to execute the design for query and if necessary, repeating the steps 2 and 3.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

It is clear that we should try to minimize the number of the steps 2 and 3. Considering step 2 includes relation with some expense our proposed algorithm has tried to minimize the restoration from a great collection of data. Our proposed algorithm searches for all probable designs for query by using 'Up to down' method and optimal rule in the least time. However, algorithm finishes the work on due time and guarantees to find the optimal design for query execution,it is possible to divide the proposed algorithm in four steps as follows:

Step 1: Catalog of all joints and possible multiple joints which may be defined as a basic relation and an intermediate relation without production Cartesian multiplication.

Step 2: Creating a proposal request for the joints and estimating step 1 to scan basic tables.

Step 3: The request cots from the proposers for the joint and scan operation. If the entrance relations are the intermediate tables, only the single joint costs is requested for each joint with supposing the entrance costs had been estimated before.

Step 4: Estimating the designs' costs and related substructures as return by dynamic programming and finding optimal design for query.

Suppose the bank with relations designs as follows:

Branch (branch_name,branch_city,assets).
Client (customer_name,customer_street, customer_city).
Loan (loan_number,branch_name,amount).
Customer (customer_name,loan_number)
Account (account_number, branch_name, balance).
Deposit (customer_name,account_number).

We may distribute tables among three sites:

Site 1: Branch.
Site 2: Customer, customer, deposit.
Site 3: Loan, account (Figure 7).

Central data lexicon includes data related to the tables in the sites and defined designs for the tables. Now suppose following query:

```
SELECT customer_name, loan_number,
amount
FROM borrower , loan
WHERE borrower.loan_number =
Loan.loan_number
AND branch_name = 'Perryridge' and
amount > 1200;
```

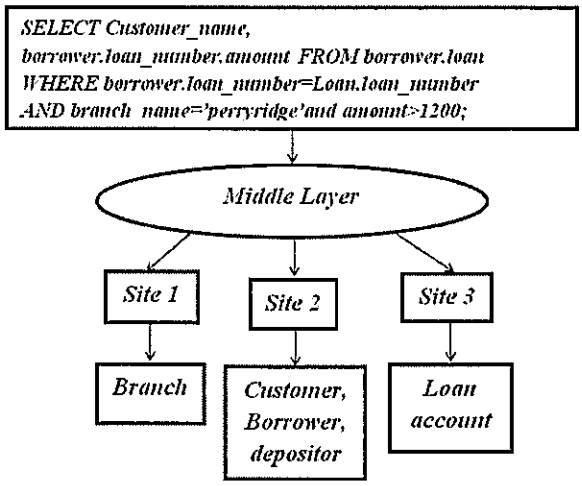


Figure 7. Proposed database

Step 1: In this step the catalog refers to central lexicon to define the sites for special query and creates 'N' SubSelects, SubWheres and fubForms in which there are 'N' sites; then SelectItems enter in SubSelect(n) and FormItems from the distributed sites in SubForms (n) (Figure 8).

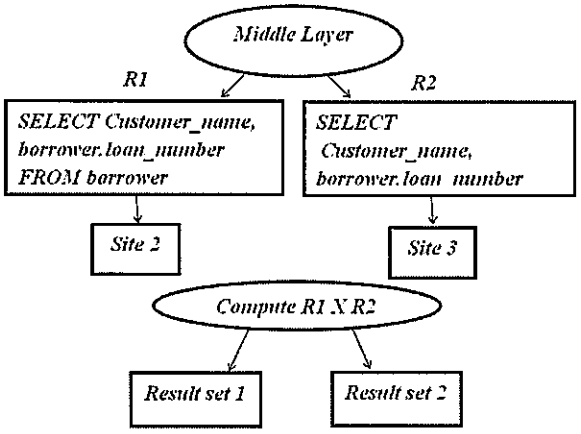


Figure 8. Step 1

Step 2: In this step we select each item from WhereItem list. If related element belongs to

that site completely, the element relates to SubWhere and if related element does not belong to a site completely, the element is located in final new list of Where (Element analysis related to the operation); then it searches the tables features one by one and finds the sites containing the features and if it does not include the features, the related SubSelect includes them. This step is one of the key steps in this algorithm; the example may describe it for us better(Figure 9).

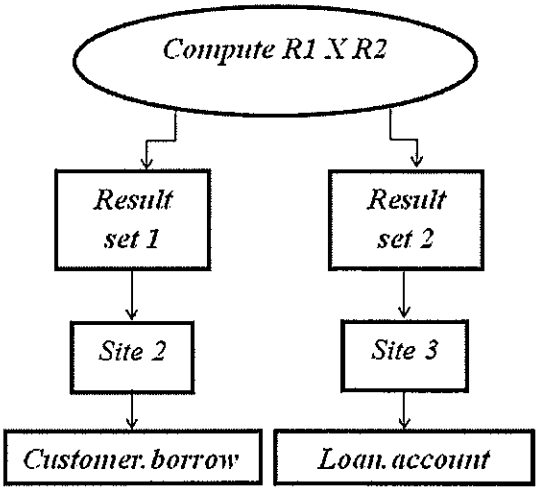


Figure 9. Step 2

Step 3: In this step we produce a SubQuery from SubSelect, SubWhere and SunForm lists.

```
FinalSelect = SelectItems
FinalFrom = {R1,R2…..,Rn}
```

Where:

n = Number of the sites.

Rn = The result of the collection achieved from 'n'; in this step we create query from FinalFrom, FinalSelect and FinalWhere lists.

Design execution:

```
SUBQUERY[1] =NIL
---------------------------------
SUBQUERY[2]:
SELECT
CUSTOMER_NAME, B.LOAN_NUMBER
FROM BORROWER ;
---------------------------------
SUBQUERY[3]:
SELECT AMOUNT, L.LOAN_NUMBER
FROM LOAN
WHERE BRANCH_NAME = 'PERRYRIDGE'
AND AMOUNT>1200 ;
---------------------------------
FINALQUERY :
SELECT
CUSTOMER_NAME,B.LOAN_NUMBER ,
AMOUNT
FROM R2, R3
WHERE
B.LOAN_NUMBER = LOAN.LOAN_NUMBER
```

Where:

R2 and R1 are the result from related parallel sites and final query in the middle layer, respectively(Figure 10).

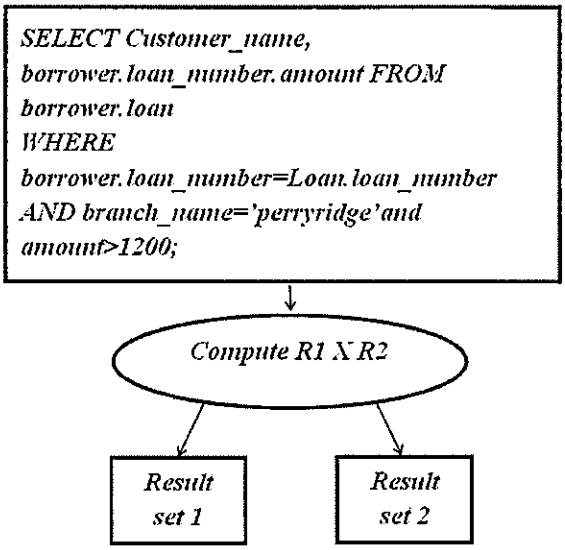


Figure 10. Design execution

Step 4: It includes queries substructure execution in related parallel sites and final query in the middle layer.

5 COMPARISON and CONCLUSION

Query optimization in distributed databank are examined from the views allocating data and producing optimal execution design, but considering both discussions are of NP there is no definitive reply for them; meanwhile, none of the presented algorithms produce optimal reply for all problems and have favorable

results only for some problems with special features; for example, the techniques querying for dynamic programming are appropriate to a little amount of queries, but such methods are not appropriate when the number of the relations in the query increases because high memory and process are used. The query optimization method for the query is very useful to distributed database systems. The optimizer should consult with the sources of the data involved in finding the operation cost to estimate the optimization process cost. In view of the cost concerning executive designs and number of the defined joints the proposed algorithm gives better results than composite evolution algorithm. The algorithms comparison indicates the proposed algorithm is better than the composite evolution algorithm. Having used this algorithm it is possible to achieve the reply more rapidly and prevent to trap algorithm in local minimums; so it can be said that the proposed algorithm is a more appropriate method to solve the problems of distributed database queries. The mentioned optimization process indicates in many cases specially when our database design is physical the query optimizer algorithm works well, but if we have not such data, we should use more aggressive optimization techniques

6 REFERENCES

[1] C. Shahabi, L. Khan, D. Mcleod. "A probe based technique to optimize join queries in distributed internet bases, Knowledge and Information Systems". Computer Science and Information Technology (ICCSIT), 3rd IEEE International Conference on, Volume 8. 2002.

[2] Z. G. Ives, D. Florescu, M. Friedman, A. Y. Levy, and D. S.Weld. "An adaptive query execution system for data integration.In SIGMOD", International Journal on Intelligent and Cooperative Information Systems,(6) 2/3:99–130, June 2009.

[3] C.Olston and J. Widom. "Offering a precisionperformance tradeoff for aggregation queries over replicated data". vldb.org/conf/2005/pp. 144.

[4] T.Oliveria ,"Evolutionary Query Optimization for Heterogeneous Distributed Database Systems", Engineering and Technology Volume33 September 2010.

[5] I. Ahmad, K. Karlapalem,and Y. Kwok, Siu-Kai So "Evolutionary Algorithms for Allocating Data in Distributed Database Systems", Distributed and Parallel Databases, January 2002, Volume 11, Issue 1, pp 5-32.

[6] S. Chaudhuri, "An Overview of Query Optimization in Relational Systems", PODS '98 Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. 1998.

[7] M. Yannis, E. Ioannidis. "Query Optimization", ACM Computing Surveys (CSUR). Volume 28 Issue 1, March 1996

[8] M. Steinbrunn, G. Moerkotte, Alfons Kemper, "Heuristic and Randomized Optimization for the Join Ordering Problem". The VLDB Journal, The International Journal on Very Large Data Bases. Volume 6 Issue 3, August 2012.

Comparison of Machine Learning Algorithms Based on Filipino-Vietnamese Speeches

Hoa T. Le
Department of Network and Communication
Thai Nguyen University of Information & Communication Technology
Thai Nguyen, Vietnam
Lthoa@ictu.edu.vn

ABSTRACT

People of different races are characterized by the language they speak. They can identify voices of someone's race just by listening and talking through conversation. This paper presents an efficient comparison of machine learning algorithm based on Filipino-Vietnamese speeches for tone classification using feature parameter. The system was trained using audio recorded speeches samples. Datasets were taken from multiple sessions involving 10 respondents; 5 (five) of which are Filipinos and 5 (five) Vietnamese. The respondents were asked to read the paragraphs and record their voices while reading the data. The empirical test shows that during the pre-processing of data records, Vietnamese have longer range of duration as compared to Filipinos because of their manners in reading and intensity on accent-bearing syllables. In constructing the speech recognition model, four classification algorithms were used, namely: KNN (K-Nearest Neighbour), Naïve-Bayes, SMO (Support Vector Machine) and MLP (Multilayer Perceptron). The evaluation of the training set in terms of accuracy, correctly classified instances and incorrectly classified instances are evaluated by the performance of the developed system. As the data established, the results show that SMO and MLP performed better for all the given datasets, with accuracy rates ranging from 99.2694% for MLP and 98.7179% for SMO. However, KNN algorithm turned out to have the lowest rate of 96.3882%.

KEY WORKS

Machine Learning Algorithm, Audio Features Parameters, Speech Recognition, Pattern Recognition, Classification Algorithms.

1INTRODUCTION

People have various races with unique languages and just by listening to someone's conversation, they can easily identify as to what racial group they belong to. The tone has conventionally been used to refer to those languages which use the feature of pitch to distinguish between lexical or grammatical meaning that is, to distinguish or to inflect words [1]. The two principle schemes for marking a tone are tonal and non-tonal languages. Conventional wisdom in automatic speech recognition asserts that pitch information is not helpful in building speech recognizers for non-tonal languages and contributes only modestly to performance in speech recognizers for tonal languages. To maintain consistency between different systems, pitches therefore are often ignored, trading the slight performance benefits for greater system uniformity or simplicity [2]. The models of tone deliver consistent performance improvements for tonal languages and even modest improvements for non-tonal languages. Using neural networks for feature integration and fusion, these models achieve significant gains throughout, and provide system uniformity and standardization across all languages, both tonal and non-tonal. The combination of multiple features for the recognition of multiple languages with different characteristics uses deep neural network [2].

Vietnamese is an Austro-Asiatic language spoken by about 82 million people mainly in Vietnam. There are also Vietnamese speakers in USA, China, Cambodia, France, Australia,

Laos, Canada and a number of other countries. Vietnamese has been the official language of Vietnam. It was originally written with a Siniform (Chinese-like) script known as Chữ-nôm(字喃) or Nôm (喃). Most Vietnamese literature was essentially Chinese in structure and vocabulary and eventually developed a more Vietnamese style, but was still full of Chinese loan words. The script is still studied and taught by some university particularly in Hanoi, which has recently published a dictionary of all the Nôm characters [3]. Roman Catholic missionaries introduced a Latin-based orthography for Vietnamese during the 17th century, Quốc Ngữ (national language) which is used until today. All morphemes which are monosyllabic compounds are joined with hyphens. The tone marks are regularly used in all writing. The six tones are indicated with the following diacritics: level (unmarked), high rising (acute), low-falling (grave), low rising or dipping-rising (superscript dot less question mark), high rising broken/gluttonized (tilde) and low-constricted/gluttonized (subscript dot) [4].

On the contrary, Tagalog is a non-tonal language with a relatively small number of phonemes. The sounds make a difference in word meaning. Tagalog has five (5) vowel phonemes; there is a contrast between short and long vowels in non-final [5]. Because the field of phonology studies sound patterns of languages, corpus-based phonology typically relies on audio corpora. The Tagalog language (Austronesia, Philippines) exhibits several morph phonological phenomena that are reflected in its spelling [6]. All of these phenomena involve some variations, which make them ideal for text-corpus study: only with large amounts of data can we investigate the distribution of the variants and search for the factors that condition the variation [6]. Zuraw (2006) presented one case study on Tagalog tapping, of phonological research using a written, web-derived corpus. Several aspects of the investigation depended crucially on the web as corpus method. Because of economic constraints, the only realistic way to assemble a large corpus of language like Tagalog is currently by taking text from the web. And only a large corpus makes it possible to ask questions such as "how does the frequency ratio of a derived word to its base affect the application of a phonological rule?" The two different patterns of variation polarized in the stem+prefix case, continuous in the word+enclitic case would have been very difficult to discover without corpus data. Tagalog corpus has already been used to investigate in fixation in loans that begin with consonant clusters. Novel method for tonal and non tonal language classification using prosodic information. Normalized feature parameters that measure the speed and level of pitch change are used to perform the classification task. By measuring the pitch changing speed and pitch changing level, this novel system can be used to perform tonal and non-tonal language [7]. Several researchers analyse and evaluate the combination of multiple features for the recognition of multiple languages with different characteristics using deep neural network. The Cantonese, Tagalog and Vietnamese systems are trained using data released within the IARPA Babel program [8]. Table 1 shows that adding tone features actually results in small gains in these languages. In Tagalog, the DBNF system with early integration reduced the WER by 1.8% compared to the baseline DBNF system. Even for the non-Babel English system, a small improvement of 0.5% from 16.0% to 15.5% could be obtained with this approach. In contrast to the tonal languages, re-initializing the system with tonal features did not result in any gain.

Table 1. Summary result obtained feature setups by merging tonal and non-tonal features.

WER/ CER (%)	ENG	TAG	CAN	VIE
Baseline	20.5%	69.0%	66.6%	68.9%
Best Baseline DBNF	16.0%	54.6%	52.8%	54.7%
Best Tonal DBNF	15.5%	52.8%	50.7%	51.7%
(rel.) over Baseline	24.4%	23.5%	23.9%	25.0%
over DBNF	3.1%	3.3%	4.0%	5.5%

In this study, the recorded voices are used to classify which algorithms are suitable for speech recognition. Four classification algorithms were used, namely KNN, Naïve-Bayes, SMO and MLP. The evaluation of training set in terms of accuracy, correctly classified instances and incorrectly classified instances evaluated the

ISBN: 978-1-941968-02-4 ©2014 SDIWC

performance of the developed system. Examining the data during pre-processing of data records show that Vietnamese have longer range of duration as compared to Filipino because of the manners in reading and linguistic form in which non-native characteristics are reflected. Section 2 discusses the related works in the field of speech recognition and language, section 3 discusses the tools, section 4 the research method, section 5 includes results and discussion and section 6 conclusions and future works.

2 RELATED WORKS

This section discusses the related works on speech languages recognition that uses different approaches. In linguistics, a phrase is a group of words (or sometimes a single word) that forms a constituent and so function as a single unit in the syntax of a sentence. In previous works, the study in [9] used two languages: English and Vietnamese. Given a phrase pe in the source language (English) and a phrase pv in the target language (Vietnamese). They defined a phrase pair p = (pe;pv) as a parallel phrase if the source phrase pe and the target phrase pv are the translation of each other, i.e., there is no additional (boundary) word in the target phrase which cannot find the corresponding word in the source phrase, and vice versa. Parallel phrases (PPs) are important for some Natural Language Processing (NLP) tasks such as machine translation or cross language information retrieval. This paper proposes a novel method to extract parallel phrases from English-Vietnamese parallel corpora. In this method, they use predefined syntactic patterns and phrase translation probabilities for determining parallel phrases. The experiments were conducted on English-Vietnamese parallel corpora and have shown that the method increases 79: 72% of F score for obtaining parallel phrases in comparison with a baseline. Index terms parallel phrase, parallel corpora, syntactic pattern, statistical measure, constrained word alignment model [9]. Any methods have been proposed for extracting PPs from parallel or comparable corpora. Methods can be classified into three main approaches: symbolic methods, statistical methods, and hybrid methods. The first approach uses a linguistic filter that depends on syntactic patterns in statistical machine translation systems; the quality of the translations is largely dependent on the quality of parallel phrase pairs extracted from bilingual corpora [10]. Actually, there are only a few studies on this task related to Vietnamese language. Nguyen [11] has proposed a method to identify base noun phrases correspondences from a pair of English-Vietnamese bilingual sentences. He detects anchor points of the base noun phrases in the English sentence, and then performs alignment relying on these anchor points. Different from Nguyen's study, this study focuses on extracting PPs from English-Vietnamese bilingual corpora. This will follow the third approach in which the researcher combine predefined syntactic patterns and phrase translation probabilities to determine PPs. Note that previous studies used syntactic patterns at two sides (both the source sentence and the target sentence) to determine PPs. Therefore, only a small quantity PPs was extracted when matching corresponding PPs in the parallel corpus. The method will overcome this drawback. Firstly, it used a set of syntactic patterns at one language to detect mono-phrases. Secondly, determined the translation of a source phrase by a constrained word alignment model and finally, the PPs with a higher probability more than a specific threshold will be extracted [9].

Another study conversed with language modeling techniques for speech recognition [12]. In this study the use of smoothing techniques are essential for n-gram-based statistical language modelling, especially in large vocabulary continuous speech recognition (LVCSR) tasks. In this study several smoothing algorithms for n-gram models in Filipino LVCSR were investigated. The automatic speech recognition system was developed using the Janus Speech Recognition Toolkit (JRTk) of Carnegie Mellon University and Karlsruhe Institute of Technology. The language models were generated using Stanford's language modelling toolkit, SRILM. The data consisted of approximately 60 hours of transcribed

ISBN: 978-1-941968-02-4 ©2014 SDIWC

recordings of Filipino speech from several domains spoken by 156 speakers. A total of 24 systems employing different language models were fine-tuned and tested for improved performance at a base metric. An instance of the Kneser-Ney algorithm with Modified-at-end counts applied to an n-gram of order 5 registered the highest word recognition accuracy at 80.9% and 81.3% for the development and evaluation tests, respectively [12].

Another study was conducted by Natural-sounding synthesized speech is the goal of HMM-based Text-to-Speech systems. Besides using context dependent tri-phone units from a large corpus speech database, many prosody features have been used in full-context labels to improve naturalness of HMM-based Vietnamese synthesizer. In the prosodic specification, tone, part-of-speech (POS) and intonation information are considered not as important as positional information. Context-dependent information includes phoneme sequence as well as prosodic information because the naturalness of synthetic speech highly depends on the prosody such as pause, tone, intonation pattern, and segmental duration. This paper intended to use decision tree questions that use context dependent tones and investigate the impact of POS and intonation tagging on the naturalness of HMM-based voice. Experimental results show that their proposed method can improve naturalness of a HMM-based Vietnamese TTS through objective evaluation and MOS test [13].

Another paper presented used the development of a closed captioning system for Filipino TV news programs are discussed. A Filipino News Corpus was built consisting of speech and text data obtained from Filipino news videos. Training and testing sets were generated and from these, different trainings and decoding parameters of Sphinx were evaluated. Using the word error rate (WER) computation, the highest average recognition accuracy achieved in developing for the test set was 57.36% using flat start context-dependent models and a language model with absolute discounting applied. This study established the baseline accuracy for future development of the system [14].

3 TOOLS USED

This paper used several tools to clean the audio file and to extract the features. Audacity software package was used to remove the unwanted signal of an audio file. jAudio software was used in feature extraction while WEKA software package was used in classifying and building a Filipino-Vietnamese recorded voices. The following are the short description of the tools used.

Figure1 presents the tool used for extracting features from audio files as well as for iteratively developing and sharing new features. jAudio extracted features can then be used in many areas of music information retrieval (MIR) research, often via processing with machine learning framework such as ACE. There are a number of aspects of jAudio that facilitate such iterative feature development. Example, jAudio uses a modular plug-in interface that avoids core code modification or recompilation when new features are added. One need only place a newly compiled feature in a plug-in folder and add a reference to it in an XML configuration file, which can refer to remote URLs as well as local file paths. Automated "met feature" extraction is another benefit of jAudio. These are template-derived features that can be extracted from one or more other feature.

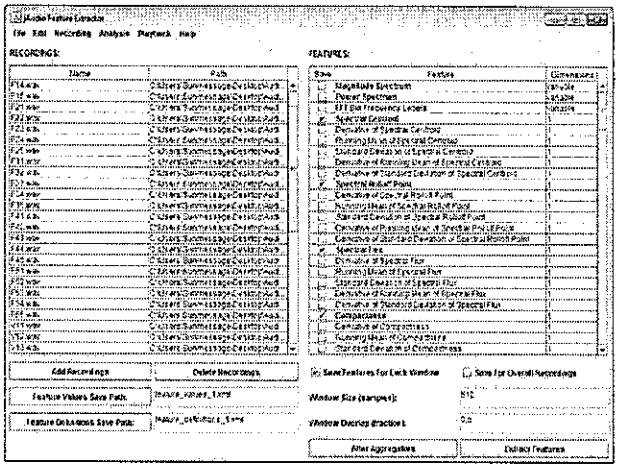


Figure 1. jAudio Extracting audio file

ISBN: 978-1-941968-02-4 ©2014 SDIWC

The paper presented by McEnnis*et. al.*, 2005 discussed the use of Audacity software that removes the unwanted signal or noise of the audio file [15]. jAudio is a new framework for feature extraction designed to eliminate the duplication of effort in calculating feature from an audio signal. It also provides a unique method of handling multidimensional features and new mechanism for dependency handling to prevent duplicated calculations.

Format Factory is a free audio, video and photo converter that supports a large range of formats for encoding and ripping. This software can convert either single files or entire folders from one format to another hence; it can easily create audio and video files that can play in different devices [16]. In this study, format factory is used to all recorded voices which convert mp4 files recorded to iPhone and convert to mp3 and wav files.

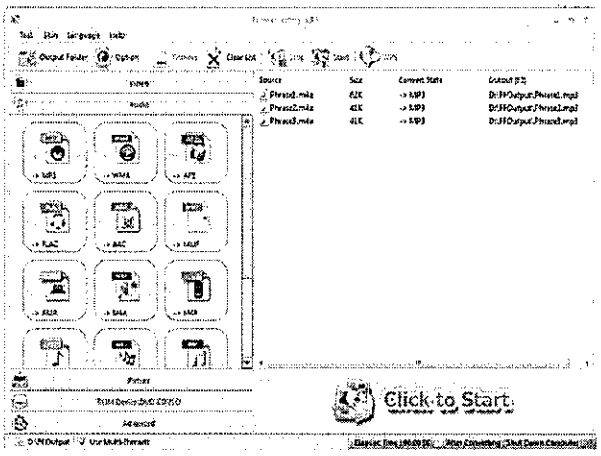


Figure 2. Format factory converter

This study conducted by Rodriguez R.L and Ilao, J.P. is entitled *Filipino Emotion Classification in Speech Signals based on Audio Features and Transcribed Text*. The Waikato Environment for Knowledge Analysis (WEKA) is being used as classifier algorithm to develop a model of Filipino emotions particularly decision tree, neural network, and support vector machine algorithm implementation [17]. In this paper, WEKA is used for classification algorithms to compare the algorithms as mentioned above for machine learning based on Filipino-Vietnamese speech.

4 RESEARCH METHOD

4.1 Experimental Setup

In this paper, all audio files are set and conducted ten (10) selected individuals, five (5) of which are Filipinos and five (5) Vietnamese. Every respondent was asked to read and record the paragraphs given. Five paragraphs were included.

"Vietnamese students should pay more attention to learning English. English can help us to keep up with people in the world."

"Learning English is a lifetime work. You are not allowed to stop learning it."

"To be good at English, please make use of review of what you have learned."

"Reading much can help you to listen better while writing help to speak smoothly."

"Vietnamese people pay much attention to reading and speaking, they do not spend enough time learning listening and writing. It is one of the weak point of Vietnamese people in learning English."

The audio files used in our study ranging from 6- 12 seconds consist of 50 audio files. The data was collected and cut each phrases and labelled the class F – Filipino and V for Vietnamese. After gathering data, pre-processing were done by using Audacity, extract features by jAudio and classifier by WEKA.

4.2 Data Set

In this study dataset has 50 records for 5 Vietnamese and 5 Filipinos; each has 5 records and used Iphone 5 to record the voices. As shown in table 2 below, the data used jAudio to extract features, the data generated with 36 features.

Table 2. Result Of Extracted Features

@relation jAudio
@ATTRIBUTE "Spectral Centroid0" NUMERIC
@ATTRIBUTE "Spectral Rolloff Point0" NUMERIC
@ATTRIBUTE "Spectral Flux0" NUMERIC
@ATTRIBUTE "Compactness0" NUMERIC
@ATTRIBUTE "Spectral Variability0" NUMERIC
@ATTRIBUTE "Root Mean Square0" NUMERIC
@ATTRIBUTE "Fraction Of Low Energy Windows0" NUMERIC

ISBN: 978-1-941968-02-4 ©2014 SDIWC

@ATTRIBUTE "Zero Crossings0" NUMERIC
@ATTRIBUTE "Strongest Beat0" NUMERIC
@ATTRIBUTE "Beat Sum0" NUMERIC
@ATTRIBUTE "Strength Of Strongest Beat0" NUMERIC
@ATTRIBUTE "LPC0" NUMERIC
@ATTRIBUTE "LPC1" NUMERIC
@ATTRIBUTE "LPC2" NUMERIC
@ATTRIBUTE "LPC3" NUMERIC
@ATTRIBUTE "LPC4" NUMERIC
@ATTRIBUTE "LPC5" NUMERIC
@ATTRIBUTE "LPC6" NUMERIC
@ATTRIBUTE "LPC7" NUMERIC
@ATTRIBUTE "LPC8" NUMERIC
@ATTRIBUTE "LPC9" NUMERIC
@ATTRIBUTE "Method of Moments0" NUMERIC
@ATTRIBUTE "Method of Moments1" NUMERIC
@ATTRIBUTE "Method of Moments2" NUMERIC
@ATTRIBUTE "Method of Moments3" NUMERIC
@ATTRIBUTE "Method of Moments4" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs0" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs1" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs2" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs3" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs4" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs5" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs6" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs7" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs8" NUMERIC
@ATTRIBUTE "Area Method of Moments of MFCCs9" NUMERIC

4.3 Machine Learning and Feature Extraction

The study used iPhone to record audio clips and Audacity to clean data. The original 50 audios still have noises or unwanted sounds. The hissing and humming sounds and clippings in the speech waveform are factors that might affect in the annotation of the respondents and might yield to the low classification accuracy result. Because of these unwanted sounds, the 50 audio files were analyzed using Audacity in order to determine which part of the audio file contains noise and to remove the unwanted sounds [18]. The noise removal algorithm uses Fourier analysis: it finds the spectrum of pure tones that make up the background noise in the quiet sound segment that are selected as discussed in the Audacity wiki [19].

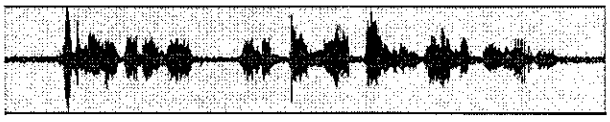


Figure 3. Audio Signal that contains noise

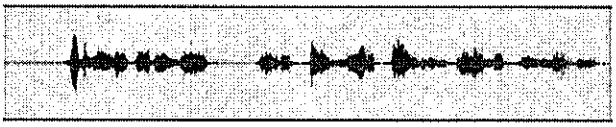


Figure 4. Cleaned Audio File

4.4 Machine Learning And Classification

In this study, two different machine learning classification algorithms from WEKA were used namely: Naïve Bayes Algorithm, KNN Algorithm, and two functional classifier, Multilayer Perceptron (MLP), SMO support vector machine implementation and clustering K-means algorithm in WEKA. The following are the short description of the classifier used in the study.

A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions [20]. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter [20]. Even if these features depend on each other or upon the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique [21].

A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If

ISBN: 978-1-941968-02-4 ©2014 SDIWC

K = 1, then the case is simply assigned to the class of its nearest neighbour.

Multilayer Perceptron is an implementation of Neural Networks algorithm in WEKA. This is the most prominent type of neural network which belongs to a class of networks called feed forward networks because they do not contain any cycles and the network's output depends only on the current input instance [17]. This algorithm is usually trained by minimizing the squared error of the network's output essentially treating it as an estimate of the class probability. A serious disadvantage of MLP that contain hidden units is that they are essentially opaque [17].

SMO is a support vector machine (SVM) implementation in WEKA. This algorithm use linear model to implement nonlinear class boundaries. SVM makes use of a separating line called hyper plane (linear model) to classify two clusters [17]. SVM are based on an algorithm that finds a special kind of linear model: the maximum margin hyper lane. The maximum margin hyper lane is the one that gives the greatest separation between the classes – it comes no closer to either than it has to [17].

The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centres to model the data; however, k-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

5 RESULTS AND DISCUSSION

Four different classification algorithms were used in classifying Filipino and Vietnamese speeches from acoustic information. KNN, Naïve Bayes, MLP and SMO were used as a classifier. These classification algorithms are WEKA implementation.

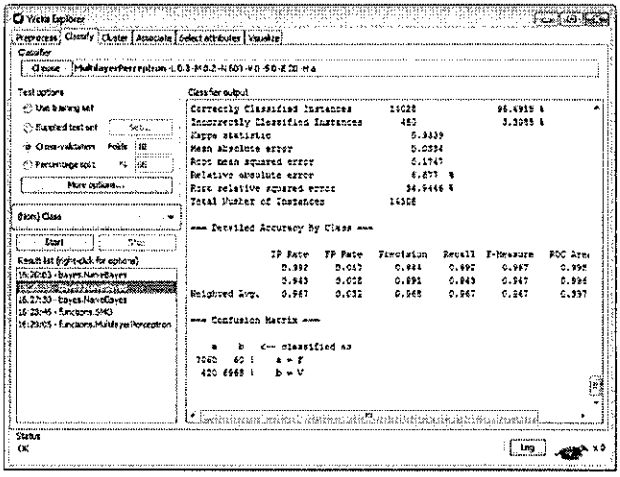


Figure 5. Naïve Bayes Algorithm

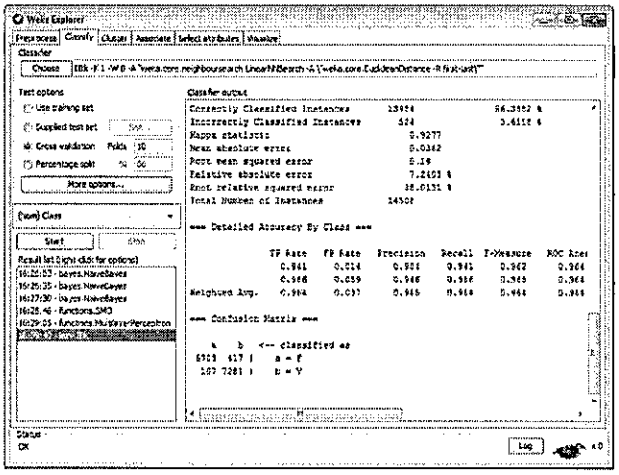


Figure 6. KNN Algorithm

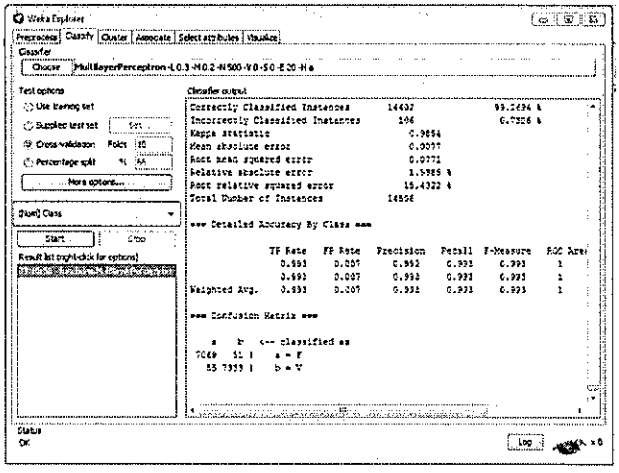


Figure 7. Multilayer Perception

ISBN: 978-1-941968-02-4 ©2014 SDIWC

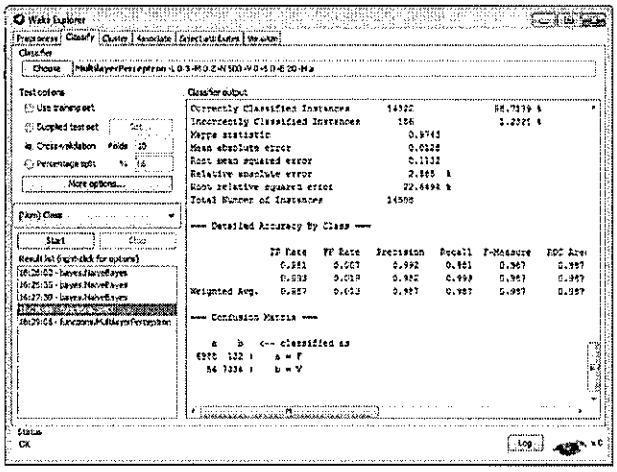


Figure 8. SVM Algorithm

Figures 5 to 8 present the performance of each model using the 10-fold cross validation and Table 3 present the performance of each model using the 10-fold cross validation.

Table 3. Result of Classification

CLASSIFIER	CCI	ICI	KAPPA STATISTICS
KNN	96.3882%	3.6118%	0.9277
NAÏVE BAYES	96.6915%	3.3085%	0.9339
SMO	98.7179%	1.2821%	0.9743
MLP	99.2694%	0.7306%	0.9854

The study is also considered as the performance measure for the speech models: correctly classified instances (CCI), and Kappa Statistics. Based on the results shown from figures 5 to 8 and as summarized in table 3, all of these algorithms got significant results. However MLP performs generally the highest performance with 99.2694% correctly classified instance with kappa statistics of 0.9854 and the second is SMO (98.7179%, 0.9743). The lowest is KNN Algorithm (96.3882%, 0.9277). Generally speaking, all models can be used to build a classifier model based on all the algorithms with high result but spend time and memory are SMO and MLP, KNN and Naïve Bayes were used in low results and slow algorithms. The results further demonstrate that all of the approaches used are good indicators to detect nationality based on speech or tonal in a conversational set-up.

This study did not consider the balancing of the dataset. As presented in the previous section of this paper, the number of audio file is equal to each category of nationality but the length of the audio file is not the same, especially those dataset that the respondents annotated (voice and word/text). Balance dataset is necessary in creating a model. The experiment set-up can be further improved, where in all annotators can listen individually to the audio file to avoid distraction and being influenced by their peers. We can generally say that the speech of Vietnamese is longer than Filipino based on the result.

6 CONCLUSIONS AND FUTURE WORK

This paper compares the speech of Filipino and Vietnamese when speaking English. The result shows that during the pre-processing of data records, Vietnamese have longer range of duration as compared to Filipinos because of their manners in reading and intensity on accent-bearing syllables.

This paper successfully presents an efficient comparison of machine learning algorithm based on Filipino-Vietnamese speeches for tone classification using feature parameter.

Consequently, the machine learning algorithms used in this study demonstrate a higher accuracy result for all dataset given. The result of this study can be further analyzed to produce good speech model that can be used in speech recognition system. However, the study is still simple with limited data and small area. The researcher would like to expand this study in other applications to indentify nationality base on their speeches.

7 REFERENCES

[1] Moria Yip (2002). Tone. (Cambridge Textbooks in Linguistics.) Cambridge: Cambridge University Press. pp 1-3, 12-14

[2] Florian Metze, Zaid A. W. Sheikh, Alex Waibel, Jonas Gehring, Kevin Kilgour, Quoc Bao Nguyen, Van Huy Nguyen. 2013. *Models of Tone for Tonal*

and Non-Tonal Languages. IEEE, Pittsburgh, PA; U.S.A.

[3] Simon Ager Omniglot - writing systems and languages of the world, 1998-2014. http://www.omniglot.com/writing/vietnamese.htm last accessed August 30, 2014.

[4] Ritter, R. M. 2002. Oxford Guide to Style: *The Style bible for all Writers, Editors, and Publisher.* Oxford University Press. New York, 2002, p. 366

[5] Irene Thompson. June 20, 2014 http://aboutworldlanguages.com/tagalog last accessed July 30, 2014.

[6] KieZuraw, 2006. Using the web as a phonological corpus: a case study from Tagalog. In *Proceedings of the 2nd International Workshop on Web as Corpus.* ACM, Stroudsburg, PA, USA, 2006, p. 59-66.

[7] Liang Wang, Eliathamby Ambikairajah and Eric H.C. Choi. "A Novel Method for Automatic Tonal and Non-Tonal Language Classification", in *Proc. of the IEEE International Conference on Multimedia and Exp* (ICME'07), China. pp. 352-355.

[8] Intelligence Advanced Research Projects Activity, "IARPA-BAA-11-02," http://www.iarpa.gov/solicitationsbabel.html, 2011, last accessed July 7, 2013.

[9] Quang Hung Le, Anh Cuong Le, and Van Nam Huynh. "Parallel Phrase Extraction from English-Vietnamese Parallel Corpora", IEEE. 2013 pp. 175.

[10] Stephan Vogel, "Pesa: Phrase pair extraction as sentence splitting," in *In* Proceedings: the tenth Machine Translation, 2005.

[11] Hieu Chi Nguyen, "A combination system for identifying base noun phrase correspondences," in Advanced Methods for Computational Collective Intelligence Studies in Computational Intelligence, Volume 457, 2013, pp. 13–23.

[12] Federico M. Ang, Juan Carlo Miguel C. Ancheta, KarmelaMariz F. Francia, and Krisel G. Chua. 2002.*Evaluation of Smoothing Techniques for Language Modeling in Automatic Filipino Speech Recognition.*EEE Institute, University of the Philippines – Diliman, Quezon City, Philippines.

[13] Thanh Son Phan, Tu Cuong Duong, Anh Tuan Dinh, Tat Thang Vu, Chi Mai Luong. 2013. *Improvement of Naturalness for an HMM-based Vietnamese Speech Synthesis using the Prosodic information.* IEEE RIVF International Conference on Computing & Communication Technologies -Research, Innovation, and Vision for the Future (RIVF).

[14] Federico Ang, MariaCzarina Burgos, and Marvin De Lara. 2008. *Automatic Speech Recognition for Closed-Captioning of Filipino News Broadcasts.* EEE Institute, University of the Philippines-Diliman Digital Signal Processing Laboratory Quezon City, Philippines.

[15] Daniel McEnnis, Cory McKay, Ichiro Fujinaga , Philippe Depalle. 2005. *jAudio: A feature extraction library.* Proceeding of the International Conference on Music Information Retrieval, 600-3.

[16] Download from: http://filehippo.com/download_format_factory/ on 27 May 2014.

[17] Ramon L. Rodriguez, Joel P. Ilao, *Filipino Emotion Classification in Speech Signals based on Audio Features and Transcribed Text.* College of Computer Studies De La Salle University, Manila Philippines.

[18] Access on 20 June 2014 from: http://audacity.sourceforge.net/about/

[19] Access on 20 June 2014 from: http://wiki.audacityteam.org/wiki/How_Noise_Removal_Works

[20] Access on 2 July 2014 from: https://www.princeton.edu/~achaney/tmve/wiki100k/docs/Naive_Bayes_classifier.html

[21] Subodh S. Bhoite, Prof. Sanjay S. Pawar, Mandar D. Sontakke, Ajay M. *Color Texture Classification Using Local & Global Method Feature Extraction* Volume No.02, Issue No. 08, August 2014 ISSN (online): 2348 – 755 p 25-31.

Estimating Tea Stock Values Using Cluster Analysis

Amitha Caldera,
University of Colombo
School of Computing
35, Reid Avenue,
Colombo 03
Sri Lanka
hac@ucsc.cmb.ac.lk

Sajitha N. Kaluarachchi
Virtusa (Pvt.) Ltd. , 752
Danister De Silva
Mawatha, Colombo 09
Sri Lanka
snksajitha@yahoo.com

Dilini T. R. Serasinghe,
Kingslake Engineering
Systems (Pvt) Ltd.
No: 30,Temple Lane
Colombo 03, Sri Lanka.
dilinitr@gmail.com

ABSTRACT

Asia Siyaka is a leading tea broker which has about 15% market share and annually sells tea with the worth of about 18 billion rupees. Tea brokers auction tea on behalf of the factory owners. Many factories buy tea from non-factory holding growers but it takes time for green leaves to become cash at the auction. Hence factories borrow money from the brokers on behalf of future stock to be auctioned to ensure seamless cash flow. The brokers' main challenge is to estimate the stock which will be sold in the future auctions before granting advances. The current system uses previous month's realized auction average prices of each factory to estimate the each factory's stock. This method is incorrect due to the variation of the grade mix of factory wise production and the price variation of tea grades with time without any pattern. Therefore requirement exists to discover an accurate method to estimate the tea stock. This paper used initial descriptive statistics which helped to understand the data and the current system. The paper also used more explorative approach because of the complexity of the problem and unclearness about the relationship of the data. Therefore cluster analysis, which is an unsupervised learning technique, was selected. The analysis shows that comparison should be done with the most recent auction data and also shows that factory, grade and package weight are the only visible attributes which can contribute to future stock value calculations.

KEYWORDS

Tea Auction, Clustering, Tea stock value, Estimation.

1 INTRODUCTION

'Tea' is an age old industry in Sri Lanka generating considerable export revenue. Majority of tea produced in Sri Lankan factories are sold via Colombo tea auction to their buyers for further processing, blending, branding and exporting.

Tea growth in Sri Lanka has a diversified distribution of estate holder statues from estate holder companies to non-factory holding growers. Average production process takes five to six weeks making green leaves auction ready. Therefore, factories tend to buy tea from small scale growers ahead of production. Time gaps between these purchases and auctions tend to block cash flow for the factories, requiring them to request advance money usually in millions of Sri Lankan rupees from the brokers based on a promised pending stock to be delivered to the broker for the upcoming tea auctions. It is inevitable for the brokers to provide these advances so that the factories in business with the broker get quality tea leaves for production at the right time resulting in better sales at the end of an auction. However, it is also a tricky business for the brokers themselves to issue advance money to a large number of factory owners within a very short period of time without being aware of the auction value of teas that is promised by the factories.

Main objective of this paper is to suggest a method of estimating a promised stock value within a limited time period. It analyses Colombo tea auction price patterns for past years in terms of weekly, monthly and annually as majority of island's tea production is sold via the auction.

Dataset for analysis was extracted for the 'low grown' elevation of teas which dominates the Colombo auction displaying 59% of auctioned production by year 2010 with respect to high

ISBN: 978-1-941968-02-4 ©2014 SDIWC

and medium elevations [1]. Also, low grown are the teas grown between the sea level and up to 600 meters. This geographical region of the country consists of many non-factory holding growers from whom medium sized factories buy teas for production, hence are frequent candidates of the loan advance process.

This further analyses auction prices of selected tea factories and tea grades. This analysis was done up to the initial level of available data. Thereafter, uses clustering technique in data mining to analyse further into attribute level.

2 PROBLEM

Determining value of a tea stock is usually done via a valuation process which consists of verifying tea samples for look and feel, smell and even taste by experienced tea tasters. However, providing a loan to the factories cannot wait for the promised stock to be evaluated. Therefore, tea brokers continuously face the challenge of issuing quick intermittent loans in millions of rupees to factories without having to do a valuation.

Currently, tea brokers consider average prices of past sales and total auctioned weight of teas of the advance requested factories before granting the loan. They look at average sales of past week, past month and past 3 months in deciding the loan value to be granted to factory. These verifications display lack of reasoning as in which averages and weights should be picked to determine the loan amounts. Also, this is more of an intuitive decision making rather than a formal one. Hence this process is vulnerable to human error and needs identifying solid relationships between attributes in order to determine which ones can be taken into account during the loan estimation process.

3 METHODOLOGY

Although data mining has already proved to be successful in many business applications, little research has been done in applying data mining techniques on Sri Lankan tea industry data. An investigation of relationships among factors such as production, export and auction price of tea using Time Series and Cluster analyses can be found in [2]. More researchers have also focused on network clustering which gather and analyze network data on an unprecedented scale to extract knowledge from networks such as social networks, biological networks, etc., [3-5]. But no formal study was found in estimating stock at Tea Auctions. The current work used actual tea auction data from 2000 to 2010 from six tea factories which are maintaining regular production patterns of 20 low grown tea grades. The initial analysis was conducted using time plots, stacked bar charts and cross tabulations to explore the dataset for relationships among production, price and time [6].

Based on the finding at the initial analysis, cluster analysis [7-12] was conducted to explore and understand the trends at Tea auctions in tea grades, package weights, elevations and factories. WEKA tool was used for the cluster analysis [13]. Clustering based on classical K-Means algorithm and SimpleKMeans clustering algorithm in WEKA is used to discover the hidden knowledge among these factors [14].

The major challenge of K-Mean clustering is to find the most optimal number of clusters and the 'Elbow finding' technique was used for the purpose [15]. In this technique, the sum of squared error is calculated for the different values of k(number of clusters) and select k value at the elbow as the optimized k as shown in Figure 10.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

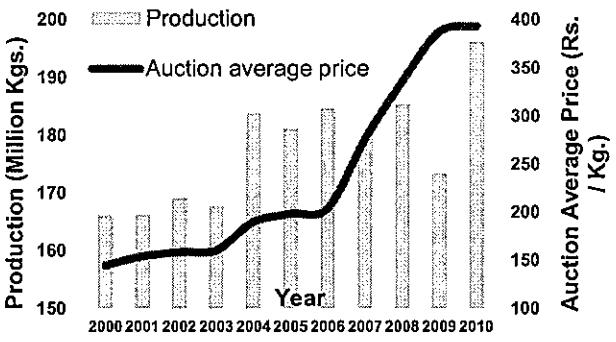


Figure1. Annual Tea production & auction average prices (2000 - 2010)

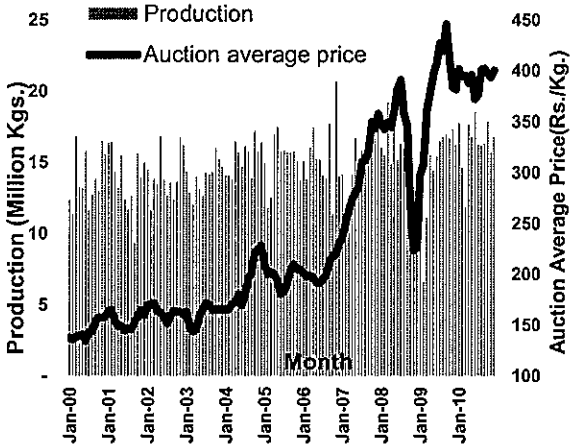


Figure 2. Monthly Tea production &auction average prices (2000 - 2010)

4 ANALYSIS

4.1 Low Grown Tea Auction Price Patterns:

As the first step, national low grown monthly and yearly auction prices and production patterns from year 2000 to 2010 were analyzed in Figure 1. According to the analysis there is no specific production pattern. But annual average price shows continuous increase. Compared to the increase from 2000 to 2006, the increase from 2006 to 2009 is very much significant. But monthly auction average prices shown in Figure 2 does not show any particular pattern. It increases as well as decreases. The sharp decrement from August 2008 to May 2009 is due to the global economic crisis [15].

The percentage difference of monthly average price to the annual average price is calculated in Table 1 which shows the effect of monthly average price to annual average price. It shows that the contribution of each month towards the annual average price is not consistent throughout the year. While some years shows positive contribution, other years shows negative. Therefore no consistent patterns for monthly average price can be found her.

The similar analysis was done between consecutive months. But as shown in Figure 3, it also shows increments as well as decrements. Therefore, no consistent patterns for monthly average prices between consecutive months can be observed here too.

Table 1. Percentage difference of monthly average price to its annual average price.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2000	-4.9	-6.0	-4.4	-3.6	-2.9	-7.1	-1.0	-0.5	6.1	8.9	8.8	9.3
2001	6.1	6.7	0.5	-3.4	-3.4	-6.1	-5.0	-5.5	-1.2	3.0	6.5	3.6
2002	6.9	7.5	8.1	3.1	2.3	0.2	-4.4	-0.2	3.3	2.7	2.6	1.2
2003	1.7	-8.3	-10.5	-9.1	-0.6	2.6	7.0	5.9	2.5	2.7	3.3	2.9
2004	-12.9	-12.8	-10.0	-9.1	-5.0	-10.6	-5.4	3.4	6.4	16.5	18.2	20.3
2005	6.3	0.9	2.0	1.0	-2.6	-9.1	-8.0	-3.4	2.8	5.8	3.3	3.3
2006	-1.2	-3.0	-2.7	-3.5	-5.9	-6.0	-4.3	-2.3	3.8	7.4	7.6	13.3
2007	15.1	-9.3	-5.0	-1.2	1.6	3.1	12.2	12.7	15.7	26.4	24.3	29.2
2008	3.4	1.9	3.9	2.6	7.6	13.7	16.3	6.3	3.1	-21.7	-33.5	-32.5
2009	-23.2	-20.6	-7.0	-2.6	2.5	5.0	10.3	8.1	15.1	6.0	-0.8	-1.5
2010	2.2	0.4	0.6	-1.3	0.8	-5.3	-4.1	2.1	2.4	1.2	0.1	1.8

ISBN: 978-1-941968-02-4 ©2014 SDIWC

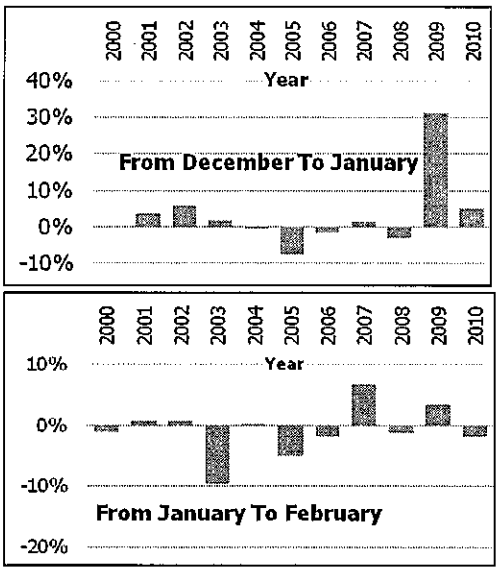


Figure 3. Percentage variance of monthly auction average price between months (2000 – 2010)

Usually there are four weekly auctions conducted per month. Therefore monthly auction average price is a combination of four auction prices. To identify the changes in monthly auction average price, the weekly low grown auction average prices are analyzed and shown in Figure 4.It also follows the same monthly pattern as in Figure 2. There is no significant variation in weekly pattern too compared to monthly pattern. Therefore the relationships between auction average prices of two consecutive weeks were further analyzed.

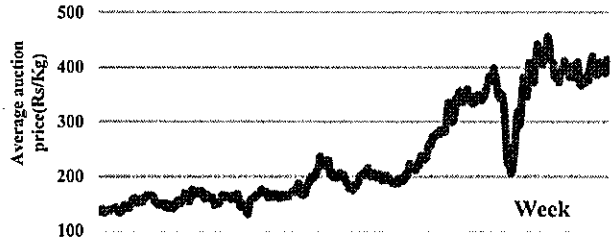


Figure 4. Weekly low grown tea auction average prices (Rs. /Kg.) (2000 - 2010)

When the unusual increments and decrements that have happened due to global economic crisis between August 2008 to May 2009 is ignored, Figure 5 and Figure 6 show just about 5% maximum of weekly difference and monthly difference respectively. This result too suggests that change of the auction price with the time is unpredictable. The most recent consecutive auctions are the closest event that can affect the price variance of tea. The results will be more accurate if the system uses auction averages of each factory of previous weeks. However, the same factory can produce the same mark of teas from different grade combinations resulting in a drastic price change from the previous auction. Therefore, grade mix differences between stocks sent to auctions can also have an impact on the results.

As the weekly auction average price is formed by a group of tea factory averages, further analysis was conducted to see any changes in weekly auction average price of each factory separately.

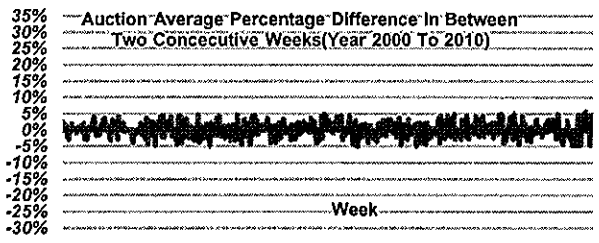


Figure 5. Auction average price percentage difference in between weeks (2000 - 2010)

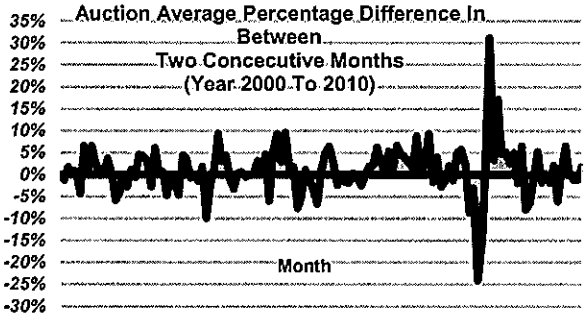


Figure 6. Auction average price percentage difference in between months (2000 - 2010

4.2 Factory Wise Auction Averages:

Monthly factory auction averages from January 2007 to December 2010 were considered in this analysis. Factories that are portrayed by the MF Codes here are some of the key low grown tea manufactures during the time of the study. As can be seen in Figure 7, the amount of increments and decrements are different from factory to factory. Further it shows that average auction prices even for a particular month varies from factory to factory.

Tea grade is the next factor that impacts on the factory average price. Therefore, further analysis was done per each grade separately.

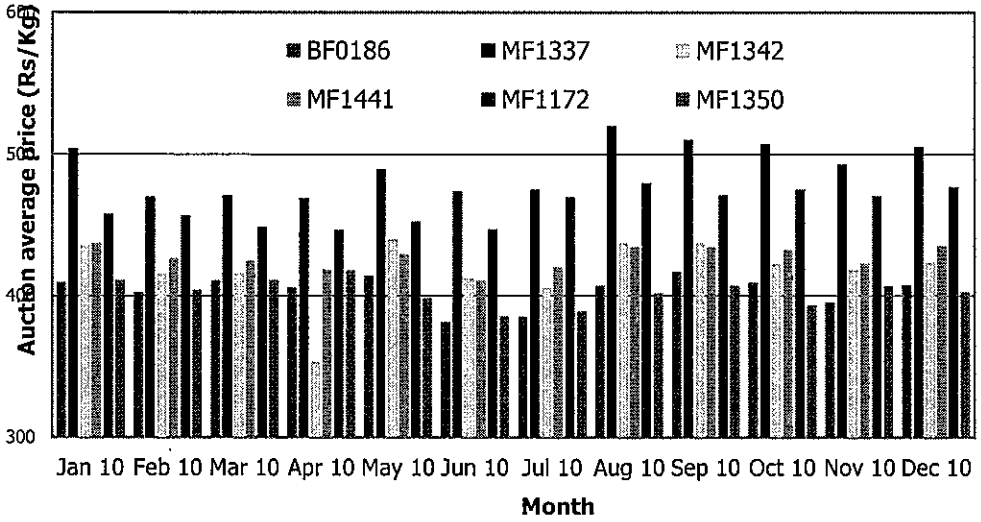


Figure 7. Factory wise monthly auction average price (Rs. /Kg.) – Year 2010

Table 2. Monthly auction quantity percentages by grade for year 2010 - MF1441

Grade	JAN	FEB	MAR	APR	MAY	JUNE	JULY	AUG	SEP	OCT	NOV	DEC
OP1	5.1	7.9	5.4	8.1	5.4	8.0	5.4	5.1	7.2	8.7	7.0	6.3
OP	6.1	7.9	9.5	10.1	8.3	6.9	9.0	7.1	9.1	6.5	4.4	6.1
OPA	5.9	7.3	5.9	9.0	9.4	7.4	9.4	6.7	7.9	6.3	10.3	6.8
BOP1	8.4	5.5	6.5	2.7	4.8	5.1	4.3	5.4	3.5	5.3	4.8	7.2
PEK	19.1	21.4	17.6	19.8	19.7	18.9	21.6	19.8	16.0	14.0	13.5	16.5
FBOP1	0.5	0.2	0.9	1.8	0.9	1.5	1.9	1.5	1.8	2.2	3.7	2.4
FBOP	13.7	13.7	15.8	9.8	13.6	13.3	11.9	11.1	12.4	13.2	11.5	14.2
FBOPF1	12.5	11.9	11.0	9.7	12.1	10.0	13.2	12.0	10.7	14.2	13.9	9.3
FBOPF	3.8	4.9	2.6	1.0	3.2	5.8	4.5	3.1	0.8	2.4	2.7	2.9
FBOPFS	2.5	2.1	0.8	2.6	2.4	2.5	2.1	3.0	1.6	3.6	3.3	2.5
FBOPFEXS	0.2	0.2	0.3	0.1	0.3	0.3	0.4	0.3	0.5	1.2	3.2	0.1
BOP	4.7	4.0	5.0	4.2	3.9	5.6	3.5	7.0	4.3	2.5	3.0	3.3
BOPF	0.5	0.4	0.0	0.0	0.9	0.0	0.0	1.4	4.0	0.0	0.0	0.0
BOPFS	0.5	0.5	1.4	0.0	0.0	0.5	0.0	0.0	1.2	4.7	4.2	4.4
BOP1A	0.7	0.0	0.4	2.6	0.6	0.8	2.4	1.3	1.9	1.5	1.5	3.2
BM	2.1	2.2	4.0	2.6	3.1	2.6	2.5	3.7	2.7	2.7	2.7	3.0
FGS	3.5	2.9	3.9	4.7	1.8	2.4	0.9	3.5	4.2	2.2	3.1	4.0
PEK1	8.1	6.2	7.4	7.2	7.5	6.7	7.2	8.1	10.3	8.8	7.2	7.9
FGS1	2.1	0.6	1.7	4.1	1.8	1.5	0.0	0.0	0.0	0.0	0.0	0.0
FBOPFEXSP1	0.0	0.0	0.1	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
TOTAL	100	100	100	100	100	100	100	100	100	100	100	100

4.3 Grade Wise Tea Auction Averages For A Factory:

The analysis of monthly auction quantity percentages by grade for a factory was shown in Table 2 above. It shows monthly auction quantity of a factory is formed by different grades. Some grades such as PEK, FBOP and FBOPF1 contribute high percentages but some grades such as FBOPFEXSP1 and FBOPFEXS contribute low percentages. Grading depends on plucking and manufacturing processes of the teas and is currently done on availability and intuitive demands for each grade. Therefore, monthly percentages of grade mix vary without a pattern. The analysis done for the auctioned prices shows the same results too.

According to the above analyses shown in Figure 7 and Table 2, there is no similarity of quantities or prices between consecutive months. Therefore there is no possibility that the sold grade percentages of the last month is similar to grade percentages of the pending tea stock in quantity or prices. In the current system the stock is estimated based on the average auction prices of previous week, month or three month. Due to the variations in grade percentages of current stock against sold stock the current method is inaccurate.

The average auction prices of individual tea grade of each factory was analyzed next. Figure 8 shows the analysis of weekly realized auction prices of tea grades, BOP1 and FBOP, at the auction during June 2010 to May 2011 for factory MF1441. These prices are in vast ranges because of their inherent quality.

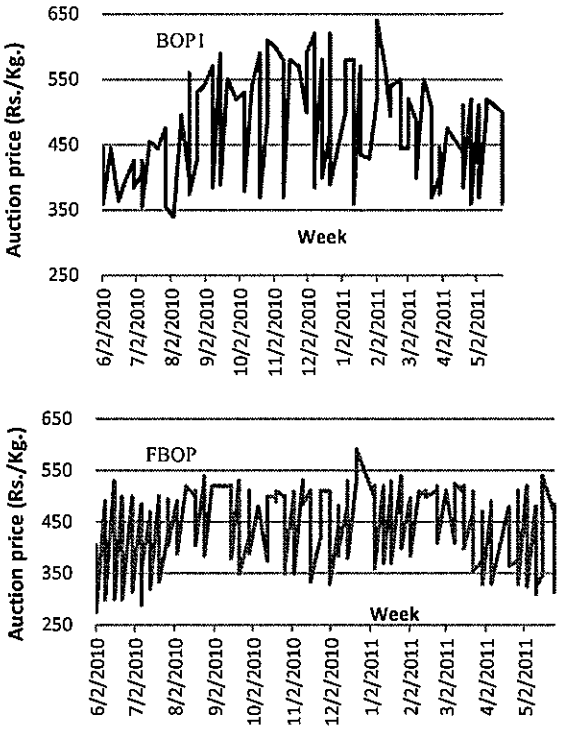
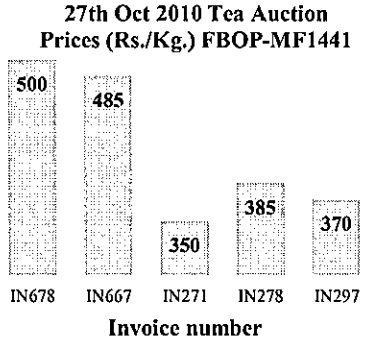


Figure 8. Weekly realize auction prices for each invoice - Factory MF1441

As shown in Figure 9, the different prices of the same grade of the same factory on the same auction are possible. This results due to the different auction conditions at different times on the same day. Therefore, the tea grades along also cannot be used as a base for stock value estimation due to the price variances in same grade in same auctions.



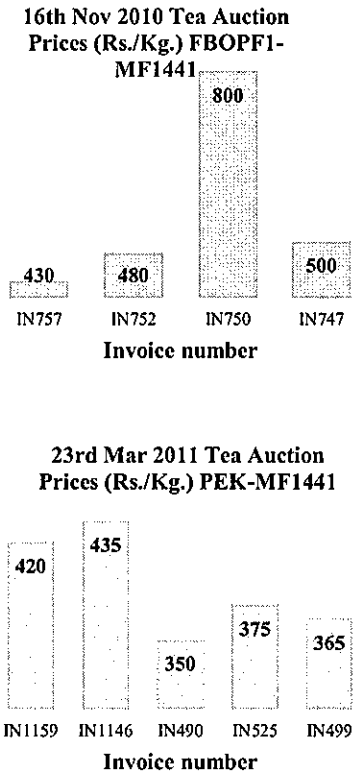


Figure 9. Invoice vs tea auction price of grades of MF1441 in three auctions

To identify the exact value of the tea invoice by invoice, a detail tasting session has to be done. But, this is not practical for stock estimation in loan granting process. Therefore, all these analysis suggest that the estimate should be done using the most recent realized auction. But it is incorrect to use realized auction average price of previous auctions of factories to estimate their stock because of the grade mix difference between stock and sold tea. Therefore, further analysis was conducted to find out a pattern within data for the different auction prices of a same grade of the same factory on the same auction. WEKA tool was used for this analysis to find the relationship between these attributes.

4.4 Grade Wise Tea Auction Averages For A Factory Using WEKA:

This analysis of grades by factory was conducted for the dataset of June 2010 to May 2011 selected from the original dataset. Using elbow finding technique, the optimized number of clusters that should be used in analysis was determined as shown in Figure 10. As abrupt change occurs at three, it is selected as the most suitable number of clusters.

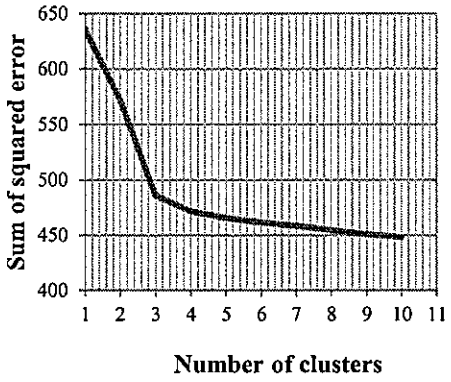


Figure 10. Number of clusters vs. Sum of squared error of FBOP of MF1172

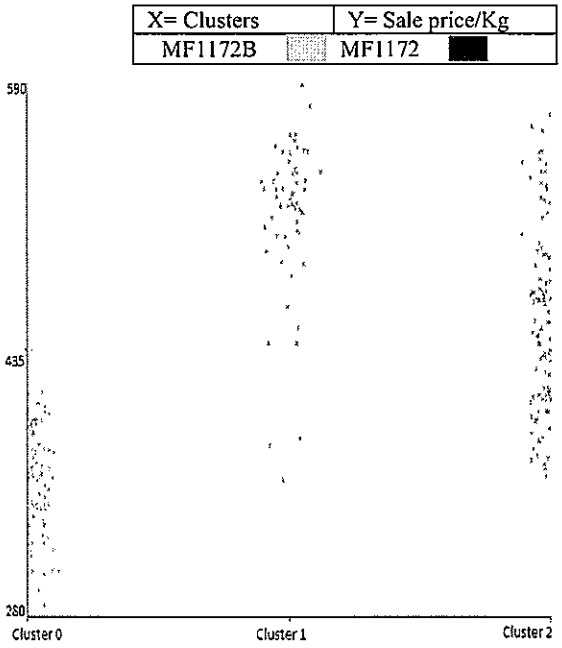


Figure 11. Clusters of FBOP of MF1172 based on price.

Figure 11 shows the result of clustered data of FBOP grade of the factory MF1172 against the price of tea. As clusters are formed under two selling marks, MF1172B and MF1172, it can be concluded that there is a relationship between the factory selling mark and the price of tea grade.

The Analysis of the same data with the package weight is shown in Figure 12. The

instances with low packing weight (32 Kg – 33Kg) were grouped into MF1172B cluster and other instances with package weight (38 Kg - 44Kg) into MF1172 clusters.

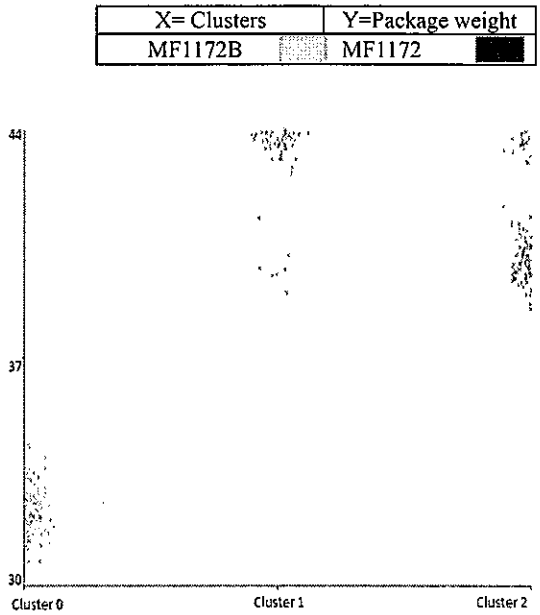


Figure 12. Clusters of FBOP of MF1172 base on package weight.

The analysis between package weight and the auction price is emphasized in Figure 13 which shows the result of package weight and Sale price of FBOP grade of the factory MF1172. Therefore this proves instances with low packing weights have low auction prices and instances with high packing weights have high prices.

The analysis in Figure 14 confirms this relation. It analyses the auction prices weekly between June 2010 and May 2011 by fixing grade, weight and factory selling marks. It shows that the auction price varies by only Rs. 100 for the entire period under one package weight. In addition to this, the price variance of a grade in between consecutive auctions is very low for the same package weight. The identified relationship can be used to calculate accurate stock value estimation.

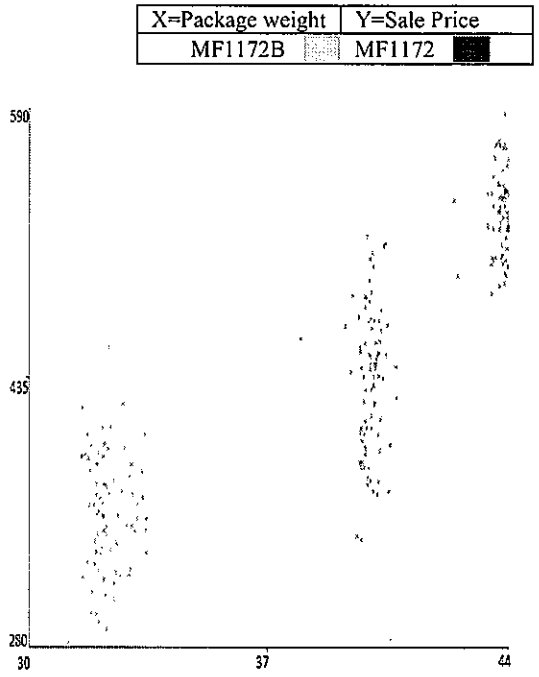
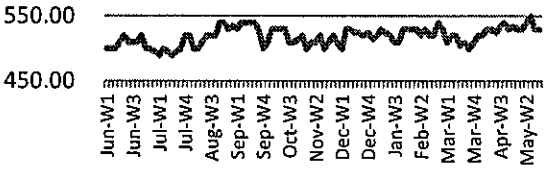


Figure 13. Clusters of Package weight vs. Sale price of FBOP of MF1172.



X Axis - Week Y Axis - Auction Price (Rs./Kg.)

Figure14. Weekly realize auction prices for each invoice for FBOPF1, MF1172, 50Kg. June 2010 to May 2011.

5 PROPOSED METHOD

According to all the analysis done and results given above, it is observed that there is a relationship between the price of a tea grade and its package weight. Even the tea with the same grade can have different package weights enforced by the Ceylon tea board. It is also observed that the realized prices for grades with same package weight of a factory at the same auction are almost equal. It is also found that the most recent realized auction prices are more suitable for the comparison of tea stock value. Therefore these relationships can be used to estimate the tea stock value.

As about three to four weeks will take to sell a tea stock, there is a three to four weeks gap between the last stock and actual

ISBN: 978-1-941968-02-4 ©2014 SDIWC

auction dates. This situation is applicable when stock date occurs within first week of the month. Worst case happen when stock date occurs within the last week of the month. Then the above gap will increase up to about seven weeks. Meanwhile previous month auction average is a combination of the respective 4 weeks auctions. It is recalled that this time gap also badly effects to the current method which is another drawback of the current system.

Combination of factory selling mark, grade and package weight are the only visible data attributes which represent the value of the grade. Therefore, these attribute can be used for the stock value estimation prior to the tea tasting. In the proposed method, when calculating stock value as at particular date, the comparison should be made between two set of invoices that are under the same tea factory, same selling mark, same grade and same package weight. That means invoice by invoice, the stock should be compared with the same type of invoices with the most recent completed auction. According to the previous analysis the auction prices of tea grades change with the time without any pattern. Therefore when doing the comparison, the most recent auction prices are to be selected. But there could be no similar grades available for comparison due to the differences in grade mix in each auction for a particular factory. If so, then it is compared with auction before the previous auction. Use the same pattern till the process finds available tea for comparison with the completed auctions.

6 RESULTS

Table 3 presents the comparison of estimated stock value by the current system and the proposed system and also it presents the actual realized value at the auction.

All instances in Table 3 proves proposed method is more accurate than the current system. The differences between actual stock value and the estimated stock value is shown in Figure 15. Red line represents difference with the current system and other line represents the difference with the proposed method.

Table 3. Estimated stock value for different dates

Factory - MF1172		Stock Value (Rs)			Deviation from Actual	
Stock Date	Stock (Kg)	Actual Auction	Current Estimate	Proposed Estimate	Current Method	Proposed Method
03Feb 2011	71,510	34,009,240	33,493,854	33,760,429	-1.5%	-0.7%
18Feb 2011	87,950	40,176,565	41,194,021	40,729,676	2.5%	1.4%
10Mar 2011	83,465	36,970,670	41,465,412	38,299,956	12.2%	3.6%
25Mar 2011	90,595	38,859,915	45,007,596	40,689,690	15.8%	4.7%
08Apr 2011	92,676	40,300,965	42,623,546	41,119,517	5.8%	2.0%
12May 2011	79,135	33,302,535	35,374,136	34,220,017	6.2%	2.8%
27May 2011	96,031	41,748,395	42,926,817	41,750,254	2.8%	0.0%
15Jun 2011	67,803	29,025,540	28,661,684	29,048,671	-1.3%	0.1%
24Jun 2011	78,978	32,562,515	31,605,929	32,398,338	-2.9%	-0.5%
05Aug 2011	87,947	33,968,120	34,636,485	33,921,966	2.0%	-0.1%
30Aug 2011	69,421	29,399,345	31,467,151	29,924,936	7.0%	1.8%

ISBN: 978-1-941968-02-4 ©2014 SDIWC

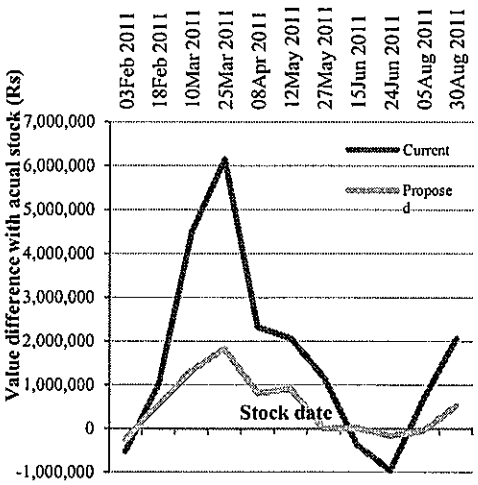


Figure 15. Stock date vs. value difference of actual and estimated

7 Future Work

There are two types of data available related to tea auction. One is data after completing the tea auction. The other one is data about three pending future auctions.

Most of the tea bought through the auction (more than 90%) is used to full fill export orders. So there is a relationship between the auction price and the pending export orders. Buyer's buying pattern in the auction depends on his pending exports orders. According to foreign client's requirement the buyer blends this tea and does a value addition before export. If there is a possibility to find pending export order details from buyers, it can improve the system further. But this data is not available in public. This is available only with the buyers. But the possibility of gathering accurate data from the buyer is un-predictable.

In addition to that the foreign currency rate change in time to time can effect to the price of tea. Therefore, the relationship between the changing patterns of the foreign currency and the rupee value can be used to improve the system further.

8 CONCLUSION

The main goal of this project is to find a solution to estimate the tea stock value using the data without tasting tea. At present the system uses auction average prices of previous month of each factory to estimate the factory stock. But after the auction it shows a huge variance between estimated price and realized price of the tea stock.

During the analysis, it was found that (1) the price of the tea varies from auction to auction, factory to factory and grade to grade, (2) the realized prices for grades with same package weight of a same factory at the same auction are almost equal and (3) the most recent auction averages are more suitable for the comparisons to estimate future stocks.

When calculating stock value for particular date, the proposed method compares two set of invoices which are belonging to the stock and the latest completed auction under the same selling mark, grade and package weight. Then the stock will be valued based on the actual realized price of the latest completed auction.

The broker's systems and databases are mainly focused and used for accounting activities. Therefore the available data with the broker is very limited for the research purpose. This was a challenge to find and organize current data for the analysis.

REFERENCES

[1]. Asia Siyaka analyses record Sri Lanka tea production of 2010 [Online] Available: http://www.ft.lk/2011/01/31/asia-siyaka-analyses-record-sri-lanka-tea-production-of-2010/

[2]. H.C. Fernando, W. M. R Tissera, and R. I. Athauda, "Gaining Insights to the Tea Industry of Sri Lanka using Data Mining", in International MultiConference of Engineers and Computer Scientists, 2008, pp. 1-6.

[3]. J. P. Bagrow, D. Brockmann, Natural emergence of clusters and bursts in network evolution, Physical Review X, 2013, vol. 3, 021016.

[4]. Y. Shang, Distinct clusterings and characteristic path lengths in dynamic small-world networks with identical limit degree distribution, Journal of Statistical Physics, 2012, vol. 149, no. 3, 505—518.

[5]. D. Q. Vu, D. R. Hunter, M. Schweinberger, Model-based clustering of large networks, Annals of Applied Statistics, 2013, vol. 7, no. 2, 1010—1039.
[6]. K.A.Smith, R.J.Wills and M. Brooks, "An Analysis of Customer Retention and Insurance Claim Patterns Using Data Mining: a Case Study", the Journal of the Operational Research Society, Vol. 51, 2000,pp. 532-541.
[7]. Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering Algorithm for Prediction of Students' Academic Performance", (IJCSIS) International Journal of Computer Science and Information Security, 2010, pp.292-295.
[8]. Ivan G. Costa, Francisco de A.T. de Carvalho and Marcílio C.P. de Souto, "Comparative Analysis of Clustering Methods for Gene Expression Time Course Data", Research Article-Genetics and Molecular Biology, 2004, pp.623-631.
[9]. V.Ilango1, Dr.R.Subramanian and Dr.V.Vasudevan, "Cluster Analysis Research Design Model, Problems, Issues, Challenges, Trends and Tools", International Journal on Computer Science and Engineering (IJCSE), 2011, pp.3064-3070.
[10]. I. G. Costa, Francisco de A.T. de Carvalho and Marcílio C.P. de Souto, "Comparative Analysis of Clustering Methods for Gene Expression Time Course Data", Research Article-Genetics and Molecular Biology, 2004, pp.623-631.
[11]. Mixotricha: The Difference between Segmentation and Clustering[Online]. Available: http://zyxo.wordpress.com/2010/07/17/the-difference-between-segmentation-and-clustering/
[12]. A. Andreescu, "Forecasting Cooperate earnings: A Data Mining Approach", Master's thesis,Swedish School of economics and Business Administration, Hanken, 2004.
[13]. Machine Learning Group at University of Waikato: WEKA [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/
[14]. Depaul University: Web Data Mining [Online]. Available: http://maya.cs.depaul.edu/classes/ect584/weka/k-means.html
[15]. R. Tibshirani, G. Walther and T .Hastie, "Estimating the Number of Clusters in a Data Set via the Gap Statistic", Royal Statistical Society, 2001, pp. 411-423.

A Method For Evaluating An Action Rule Specified By A User

Seunghyun Im
Computer Science Department, University of Pittsburgh at Johnstown
450 Schoolhouse Rd, Johnstown, PA 15904
sim@pitt.edu

ABSTRACT

An action rule is a rule extracted from a decision table that describes the expected change in the decision value of an object as a result of the changes made to some of its condition values. We use it primarily to recommend ways to improve the current state of an object. e.g., improve a person's health, student retention rate, or manufacturing process. Various methods have been proposed to automatically generate concise and effective action rules. However, a practical problem not discussed in the existing literature is the method for evaluating an action rule specified by a user. This is an important task because, in many cases, users (e.g. the decision maker of an organization) have some knowledge about specific actions they can take, and need to know the performance of the action rule to make a better decision. In this paper, we present a taxonomy of the types of action rules submitted by a user and a method for measuring the performance of the rules.

KEYWORDS

Data Mining, Rough Set, Action Rule, Rule Evaluation, Recommendation System.

1. Introduction

An action rule [1] is represented in the form of *if* $(t_1 \Rightarrow t_2)$ *then* $(d_1 \Rightarrow d_2)$, where t_1 and t_2 are the antecedents and d_1 and d_2 are the consequents of two classification rules. The rule describes the changes required in the antecedents in order to achieve the desired change in the consequent. The term *action* refers to $(t_1 \Rightarrow t_2)$ that is understood as the act of changing one or more attribute values (t_1) of an object to t_2. We use an action rule primarily to recommend ways to improve the current state of an object. i.e., improve a person's health, student retention rate, or manufacturing process. An action rule is usually built from a pair of classification rules extracted from a decision system. To illustrate the process, let us use a simple dataset in Table 1. It contains the data about the exam results of 20 students who used two types of books to prepare for an exam. The classification rules extracted from conventional data mining techniques (e.g. association rule mining, decision tree, Naive Bayes, and etc. [13]) are the frequent or interesting pattern hidden in the dataset. We use them to create an if-then type rule. For example, the classification rules from Table 1 would be *if* $book_1$ *then bad* or *if* $book_2$ *then good.* That is, the students who used $book_1$ did not do well on the test while the students who used $book_2$ did well on the test.

Table 1. Test Result of Students

	condition value	decision value
Students	Reference Book	Test Result
1 ~ 10	$book_1$	*bad*
11~ 20	$book_2$	*good*

We go one step further to find out ways to improve the student performance. The concept behind the action rule is that some attribute values of an object can be changed and controlled. We compare the classification rules to find out the attribute values that can be changed. These possible changes are used to generate an action rule that can help the students perform better on the test. (e.g.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Recommend students who used $book_1$ to change it to $book_2$).

We used a very simple example to illustrate the use of an action rule. In practice, the action rule mining task requires a more sophisticated analysis because we cannot change some attribute values (e.g. gender) or need to change more than one attribute value.

A number of methods have been proposed for generating action rules [11]. These methods put emphasis on the *automated discovery* of the action rules that the antecedent set is as *small* as possible. The idea behind the 'automated discovery' is to generate action rules from the most frequent pair of patterns, based on the assumption that users are only interested in the most significant rules. The 'small' antecedent set is motivated by the belief that a small number of elements require less changes in the objects, which would allow users to take an action with less effort. Although these methods are very useful, we have observed that users (e.g. the decision maker of an organization) have some knowledge about specific actions they can take, and need to know the performance of the action rule to make a better decision. The action may not be the shortest or the most significant action, but it is the one that can be actually performed. The existing algorithms are not designed to assess an action rule specified by a user.

This paper focuses on the case where a user provides some information about an action. We present a taxonomy of the types of action rules submitted by a user and a method for measuring the performance of the rules. The performance index proposed in this paper provides more detailed information about an action.

2. Decision System and Action Rule

We will use the following notations. By an information system S [12], we mean a triple $S=(X,A,V)$, where

$X = \{x_1, x_2, \ldots, x_i\}$ *is a finite set of objects,*
$A = \{a_1, a_2, \ldots, a_j\}$ *is a finite set of attributes,*
$V = \{a_1(x_1), a_2(x_2), \ldots, a_j(x_i)\}$ *is set of their values*

Table 2. A Decision (Information) System

	condition			decision
	flexible		stable	
Object	E	F	G	D
x_1	e_1	f_1	g_1	d_1
x_2	e_1	f_1	g_1	d_1
x_3	e_1	f_1	g_2	d_1
x_4	e_1	f_1	g_3	d_1
x_5	e_1	f_3	g_1	d_1
x_6	e_2	f_2	g_1	d_2
x_7	e_2	f_2	g_1	d_2
x_8	e_2	f_2	g_2	d_2
x_9	e_2	f_2	g_2	d_2
x_{10}	e_1	f_2	g_1	d_2

For example, we have 10 objects, $\{x_1, \ldots, x_{10}\}$, and 4 attributes, $A = \{E, F, G, D\}$, in Table 2. The value of the attribute E in x_1 is e_1, and it is written as $E(x_1) = e_1$. An information system is called a decision system if $S = (X, A_C \cup A_D, V)$, where A_C is the set of condition attributes and A_D is the distinguished attribute called the decision. A decision system is a special kind of an information system often used to extract supervised classification rules. We assume that A_C is further partitioned into stable attributes and flexible attributes. An attribute is stable if the values assigned to an object is not modifiable. Otherwise, it becomes a flexible attribute. A birth date is an example of a stable attribute. An interest rate is a flexible attribute. These attributes are denoted as,

A_{CS} = *stable condition attribute*
A_{CF} = *flexible condition attribute*

In Table 2, $A_C = \{E,F,G\}$, $A_{CF} = \{E,F\}$, $A_{CS} = \{G\}$, and $A_D = \{D\}$. We define two sets of attribute values denoted as t_m and t_n. t_m is a subset of the values of A_{CF} in x, and t_n is the set of attribute values that t_m is converted to. t_m and

ISBN: 978-1-941968-02-4 ©2014 SDIWC

t_n should be the values of the same attribute set. In other words,

t_m = the source attribute values
t_n = the target attribute values

For simplicity of presentation, we use a decision system that has two decision values, $A_D=\{d_1, d_2\}$. The objects in S can be partitioned into two groups based on the elements in A_D.

$X_{d_1} = \{x_i \in X; D(x_i) \in d_1\}$; i.e., objects that the decision values are in d_1
$X_{d_2} = \{x_j \in X; D(x_j) \in d_2\}$; i.e., objects that the decision values are in d_2.

We assume that,

d_1 = negative result
d_2 = positive result

An action rule is defined as the form of,

$$if\ (t_m \Rightarrow t_n)\ then\ (d_1 \Rightarrow d_2)$$

where $(t_m \Rightarrow t_n)$ is the possible transition of one or more attribute values of an object, and $(d_1 \Rightarrow d_2)$ is the expected change as a result of $(t_m \Rightarrow t_n)$.

Since the system has only 2 decision values, we assume the transition of the decision value occurs from d_1 to d_2. However, the presented algorithm directly extends to the general case where $|V_D| \geq 2$ by running the algorithm repeatedly for each pair of decision values.

3. Method Description

3.1 Types of User Query

Suppose that a user submits a query to our rule evaluation system in order to measure the performance of an action rule. The form of the query varies depending on the amount of information the action rule has. We can classify a user's query into two types (See Figure 1).

Type		condition		
1		no data		
2	2.1	t_m	→	t_n
	2.2	t_m	→	
	2.3		→	t_n
	2.4	A_L	→	A_L

Figure 1. Types of action rule query submitted by a user. A_L denotes a subset of A_{CF}.

Type 1. Queries of this type do not include specific attribute names or values to be used for an action. An action rule discovery algorithm will automatically generate a list of the shortest action rules that satisfy the given threshold values (e.g. minimum support and confidence values for classification rules). This is the way that action rules are extracted by the existing methods.

Type 2. This is the type of queries we focus on in this paper. A query can contain partial or full information about the action that the user is interested in. The partial information can be a set of attribute names in A_{CF} or one side of $t_m \Rightarrow t_n$. This type of queries are further divided into,

Type 2.1. $t_m \Rightarrow t_n$: Full information about an action is given. That is, both target and source values are specified. The system measures the performance of the submitted rule using the method described in the next section.

Type 2.2. $t_m \Rightarrow [\ ]$: The target values are not specified. The system needs to create a list of unique t_n s from X_{d2} (e.g. $e_2 \cdot f_2$, $e_1 \cdot f_1$ in Table 3), and measures the performance of the rules generated by each pair of t_m and t_n.

Type 2.3. $[\ ] \Rightarrow t_n$: This is similar to Type 2.2 except that we need to create a list of terms from X_{d1}.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Type 2.4. $A_L \Rightarrow A_L : A_L$ denotes a subset of A_{CF}. This type of queries contain only a set of attribute names. The system creates two sets of t_ms and t_ns in A_L and measures the performance of the action rule for each pair of t_m and t_n.

The action rules in Type 2.2, 2.3 and 2.4 can be evaluated by running the algorithm for Type 2.1 repeatedly. We will explain how to calculate the performance of the action rule of Type 2.1 in the next section.

Table 3. Finding the objects supporting *if* $(e_1 f_1 \Rightarrow e_2 f_2)$ *then* $(d_1 \Rightarrow d_2)$

	condition			decision
	flexible		stable	
Object	E	F	G	D
x_1	e_1	f_1	g_1	d_1
x_2	e_1	f_1	g_1	d_1
x_3	e_1	f_1	g_2	d_1
x_4	e_1	f_1	g_3	d_1
x_5	e_2	f_3	g_1	d_1
x_6	e_2	f_2	g_1	d_2
x_7	e_2	f_2	g_1	d_2
x_8	e_2	f_2	g_1	d_2
x_9	e_2	f_2	g_2	d_2
x_{10}	e_1	f_1	g_1	d_2

```
def find_actionable (tm, tn, d1, d2):
  Xd1, Xd2 = partition X by d1, d2
  Xtm = all x in Xd1 where x contains tm
  Xtn = all x in Xd2 where x contains tn
  Ttm = partition Xtm by stable values
  Ttn = partition Xtn by stable values
  for each t1, t2 in Ttm , Ttn
   if ((stable v in t1)=(stable v in t1))
       tm→tn is valid
       include it to the calculate of
       rule performance
```

Figure 2. Pseudo code for finding actionable objects

3.2 Action Rule Evaluation

We use Table 3 to describe our evaluation method. Suppose that a user wants to calculate the performance of the action rule, *if* $(e_1 \cdot f_1 \Rightarrow e_2 \cdot f_2)$ *then* $(d_1 \Rightarrow d_2)$. First, we need to find out the objects that can be used to build the action rule. Two objects needs to satisfy the following three conditions in order to be valid objects.

(1) different decision values (d_1, d_2)
(2) the same stable value
(3) $(e_1 \cdot f_1)$ and $(e_2 \cdot f_2)$

The reason for (1) is clear. We need to look for the difference between the objects having different decision values. (2) Stable values cannot be changed. Therefore, two objects are not counted as a valid action if they have different stable values. (3) One object in X_{d2} must have $(e_1 \cdot f_1)$ and the other object in X_{d2} needs to have $(e_2 \cdot f_2)$ in order for a transition to take place.

Figure 2 is the pseudo code for the above steps. We first partition the table into two groups based on d_1 and d_2. Then, we have

$X_{d_1} = \{x_1, x_2, x_3, x_4, x_5\}$
$X_{d_2} = \{x_6, x_7, x_8, x_9, x_{10}\}$

Next, we find the objects in X_{d_1} that contains $e_1 \cdot f_1$, and another set of objects in X_{d_2} that contains $e_2 \cdot f_2$. They are denoted as,

$X_{t_m} = \{x_1, x_2, x_3, x_4\}$
$X_{t_n} = \{x_6, x_7, x_8, x_9\}$

We group these objects by stable attributes.

$T_{t_m} = \{\{x_1, x_2\}, \{x_3\}, \{x_4\}\}$
$T_{t_n} = \{\{x_6, x_7, x_8\}, \{x_9\}\}$

The Cartesian products of T_{t_m} and T_{t_n} shows that there are two possible transitions. $\{x_1, x_2\} \Rightarrow \{x_6, x_7, x_8\}$ and $\{x_3\} \Rightarrow \{x_9\}$. The following shows the objects and their values involved in the transitions.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

$e_1 \cdot f_1 \cdot g_1 \cdot d_1 (x_1, x_2) \Rightarrow e_2 \cdot f_2 \cdot g_1 \cdot d_2 (x_6, x_7, x_8)$
$e_1 \cdot f_1 \cdot g_2 \cdot g_1 (x_3) \Rightarrow e_2 \cdot f_2 \cdot g_2 \cdot g_2 (x_9)$

We need to measure the performance of the action rule. Existing action rule mining algorithms use the concept of support and confidence [11], which is a common way to measure the performance of a classification rule. Although they are fundamentally useful, the way that the frequency of a specific pattern is counted for an action rule is different from that of conventional classification rules. For example, the frequency (or support) value 5 of a classification rule means there are 5 objects having the same pattern. As shown in Figure 3, an action rule has two of those. One from the source and the other from the target. We propose to use the following notation to measure the performance of an action rule. *Action Rule Performance* (p) is defined as the ratio between (the number of possible transitions from $e_1 f_1 \rightarrow d_1$ to $e_2 f_2 \rightarrow d_2$) and (the number possible transitions from $e_1 f_1$ to $e_2 f_2$). Let $sup(v)$ be the number of objects that has v, and $\Rightarrow$ denote a possible transition. Then,

$$p = \frac{sup(t_m \cdot d_1) \Rightarrow sup(t_n \cdot d_2)}{sup(t_m) \Rightarrow sup(n)}$$

In our example, $t_m = e_1 \cdot f_1$ and $t_n = e_2 \cdot f_2$. Therefore, p is,

$$p = \frac{sup(e_1 \cdot f_1 \cdot d_1) \Rightarrow sup(e_1 \cdot f_1 \cdot d_1)}{sup(e_1 \cdot f_1) \Rightarrow sup(e_1 \cdot f_1)} = \frac{3 \Rightarrow 4}{5 \Rightarrow 4}$$

$sup(e_1 \cdot f_1) = 5$ and $sup(e_2 \cdot f_2) = 4$ because 5 objects have $e_1 \cdot f_1$ and 4 objects have $e_2 \cdot f_2$ in S. $sup(e_1 \cdot f_1 \cdot d_1) = 3$ and $sup(e_1 \cdot f_1 \cdot d_1) = 4$ because there are 3 objects ($\{x_1, x_2\}, \{x_3\}$) in X_{d1} and 4 objects ($\{x_6, x_7, x_8\}, \{x_9\}$) in X_{d1} that a transition can occur without modifying a stable value.

Finally, the action rule is written as,

$$if\ [(e_1 \cdot f_1) \Rightarrow (e_2 \cdot f_2)]\ then\ (d_1 \Rightarrow d_2), p = \frac{3 \Rightarrow 4}{5 \Rightarrow 4}$$

It is possible to use the sum or product of two support values for p (e.g. $3 \times 4 = 12$, instead of $3 \Rightarrow 4$). However, that will obscure the quality of the possible transition because they do not show the actual number of objects involved in the transition. For instance, if there was only 1 $e_1 f_1$ and 99 $e_2 f_2$ in S, the transition would not be as significant as the transition between 50 and 50. The informative index used in this paper will help users make a better decision.

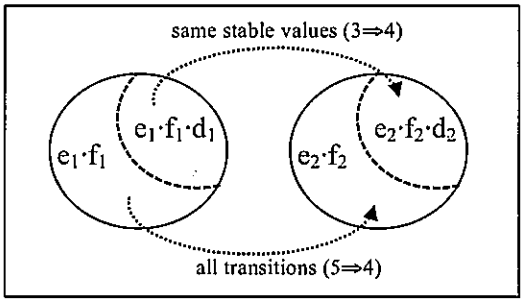


Figure 3. Measuring the performance of *if* $(e_1 f_1 \Rightarrow e_2 f_2)$ *then* $(d_1 \Rightarrow d_2)$

4. Related Work

The method for formulating an action from a decision system has been discussed in many literatures. [1] introduced the concept of the action rule discovery from a pair of classification rules. A similar concept known as interventions was discussed in [2]. The action rules in [1] has been developed further. In [3], the authors present a method for constructing action rules from a decision tree. [4] presented an algorithm for discovering action rules directly from a decision system without pre-existing classification rules. The authors in [5] also proposed a method for identifying action rules directly from a dataset. The method proposed in [6] is designed to generate action rules from a dataset containing numerical attributes. The method in [7] extracts action rules from an incomplete information system that allows an object to have many possible

ISBN: 978-1-941968-02-4 ©2014 SDIWC

values for an attribute. More recently, the authors in [8] presented an action rule mining technique that does not require an input parameter for a rule extraction and produces a consistent rule set. [9] proposed various ways of measuring the importance of an action rule. Action rules have been applied to the analysis of a wide range of data types such as medical data [10].

5. Conclusion

In this paper, we presented a method for measuring the performance of an action rule specified by a user. Our method is designed to be used when users have some knowledge about specific actions they can take. The performance index for an action rule proposed in this paper is more informative than that of the existing methods because it shows the actual number of objects involved in transition from one state to another.

6. References

[1] Ras, Z.W., Wieczorkowska, A.: Action-Rules: How to Increase Profit of a Company. In: Principles of Data Mining and Knowledge Discovery, PADD 2000. LNCS 1920, pp. 587-592. (2000)

[2] Greco, S., Matarazzo, B., Pappalardo, N., Slowinski, R.: Measuring Expected Effects of Interventions based on Decision Rules, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 17, No. 1-2, pp. 103-118 (2005)

[3] Tsay L., Ras Z.: Action rules discovery: system DEAR2, method and experiments, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 17 No. 1-2, pp. 119-128 (2005)

[4] Ras Z., Dardzinska A., Action Rules Discovery without pre-existing classification rules, In: Proceedings of RSCTC 2008 Conference, LNAI 5306, pp. 181-190 (2008)

[5] He, Z., Xu, X., Deng, S., Ma, R.: Mining Action Rules from Scratch, Expert Systems with Applications, Vol. 29, No. 3, pp. 691-699 (2005)

[6] Ras, Z.W., Dardzinska, A.: Action Rules Discovery based on Tree Classifiers and Meta-Actions. Foundations of Intelligent Systems, Proceedings of ISMIS 2009. LNAI 5722, pp. 66-75 (2009)

[7] Im, S., Ras Z.W., Wasyluk, H.: Action Rule Discovery from Incomplete Data. Knowledge and Information Systems. Vol. 25, No. 1, pp. 21-33 (2010)

[8] Im, S., Ras, Z., Tsay, L.: Action Reducts. In: Proceedings of ISMIS 2011, LNCS 6804, pp. 62-69 (2011)

[9] Tsay L.: Interestingness Measures for Actionable Patterns. In: Proceedings of RSEISP 2014, LNCS 8537, pp. 277-284 (2014)

[10] Touati, H., Ras, Z.W., Studnicki, J., Wieczorkowska, A.: Mining Surgical Meta-Actions Effects with Variable Diagnoses Number, In: Proceedings of ISMIS 2014, LNAI 8502, pp 254-263 (2014)

[11] Ras, Z., Dardzinska A.: From Data to Classification Rules and Actions, The Special Issue on Rough Sets Theory and Applications, International Journal of Intelligent Systems, Wiley, Vol. 26, Issue 6, pp 572-590 (2011)

[12] Pawlak, Z.: Rough Sets-Theoretical Aspects of Reasoning about Data, Kluwer, Dordrecht. (1991)

[13] Tan, P., Steinbach, M. and Kumar, V., Introduction to data mining, 1st edition, Pearson Addison Wesley, Boston, (2005)

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Feel the Heat: Emotion Detection in Arabic Social Media Content

Omneya Rabie
Mentor Graphics Corporation
78 Elnozha St., Heliopolis, 11361 Cairo, Egypt
omneya_rabie@mentor.com

Christian Sturm
Hamm-Lippstadt University of Applied Sciences
Maker Allee 76-78, 59063 Hamm, Germany
Christian.sturm@hshl.de

ABSTRACT

The automatic detection of emotions in textual parts of social media websites such as Facebook and Twitter has applications for business development, user interface design, content creation, emergency response, among others. Current research has shown that it is possible to detect emotions for English content. To our knowledge, however, there are only few attempts for Arabic content. There is neither Arabic corpus with instances labeled for emotions, nor studies to detect emotions from Arabic microblogs content. Therefore, we collected Arabic text messages from the social networking website Twitter from January/February 2011. Human annotators labeled them with the corresponding emotions. Working with that corpus, our experiments show that emotions can be automatically detected from tweets after performing Arabic language related language preprocessing steps. Our contribution consists in adding preprocessing steps that have improved the classification results by 4.4% compared to the original Khoja stemmer. In addition, we have extracted a sample word-emotion lexicon from that corpus. Our experiment demonstrates that this sample word-emotion lexicon enhances the emotion detection results by 22.27% compared to the SMO classification using the train/test option. Finally, we show that the communication style used by the writer significantly relates with the emotion expressed in the text.

KEYWORDS

Emotion detection, Social media, Twitter, Arabic text, Classification.

1 Introduction

Long before awareness of the World Wide Web became wide-spread, people used to seek their friends' opinions or to consult consumer reports about products or services that they want to buy. Moreover, they would make verbal surveys in order to know which candidate most of the people are planning to vote for in a local election. Lately, the Internet and the Web have made it easier to collect such data. Text is widely used in the communication between people on the web. It delivers informative content, one's opinion, and emotional state. Microblogs allow a vast pool of people that are neither personal acquaintances nor well-known professional critics to provide their experiences, opinions, and emotions.

Studies about emotions have been conducted by psychologists and behavior scientists for long time [1] [2] [3] [4]. They consider emotions a major element of the human nature. Many sentiment analysis studies have been conducted to annotate text as positive/ negative valence. However, emotion annotation can be more effective and accurate. Despite the evolving importance and usage of microblogs, there only exist few trials for building microblog corpus for emotions labeling.

Although the increasing usage of Arabic language in blogs and social networks, it is not given enough focus in the sentiment analysis or emotion detection fields. According to [5], the number of Internet users in the Middle East was 90,000,000 in 2012. In addition, the Arabic language comes at the seventh place between

ISBN: 978-1-941968-02-4 ©2014 SDIWC

the ten most used languages on the Internet. It is expected to take the fifth place in 2015. The Arabic content is in continuous increase in social media.

We are contributing to the research area of emotion detection in Arabic text by describing the experiments concerned with emotion analysis of Arabic microblogs content. The main idea is to automatically detect the emotions of microbloggers who use Arabic language. For this purpose, an Arabic corpus was built out of microblog text. Different classification techniques were evaluated. Moreover, a set of features was checked through our experiments. The aim of this project is to highlight the importance of the Arabic social media content analysis and going beyond polarity classification of text, to conduct automatic emotion analysis based on the most basic emotions suggested in [6] and [2].

This paper is structured as follows: section **2** includes some related work in emotions and sentiment analysis fields. Then, section **3** describes our experimental setup and methods for analyzing Arabic social media content gathered from Twitter. Section **4** contains the results of the different classification algorithms and the comparison between several preprocessing techniques. Finally, we conclude with a discussion of the implications that this work has for the automatic emotion detection task.

2 Related Work

"Sentiment Analysis is the computational treatment of opinion and subjectivity in text." [7]. It is the application of natural language processing and text analytics to determine the target information from the text. It is used to determine the attitude of the writer, including his/her evaluation of the topic as well as his/her emotional state with respect to some topic.

2.1 Sentiment Analysis of English Language

In the field of sentiment analysis, the work related to text polarity can be divided into two groups; **techniques to automatically generate sentiment lexicons, and systems that analyze sentiment in text documents**. The first group depends widely on adjectives. The lexicon construction starts with small seed lists and an assumption by the researcher. The hypothesis by Hatzivassiloglou and Mckeown in [8] was that if adjectives are separated by "and", they are of the same polarity; however if they are separated by "but", they are adjectives of opposite polarity. Later on, the gradation categorization was taken into consideration as well as polarity using statistical models [9]. WordNet lists [10] have been also used as seeds in the process of generation of positive and negative word lists. The assumption is that synonyms of a certain word have the same polarity, while antonyms have opposite polarity [11].

The second group is focused on **sentiment analysis systems**. The sentiment analysis of movie reviews in [12] was performed by a machine learning technique that achieved an accuracy of 83% for polarity classification. It was intended to consider subjective character to enhance sentiment analysis. However, in [11] the news sentiment analysis system considers facts and opinions as contributors to the public sentiment. The system focuses on local sentiments as more reliable than global document sentiment. The same approach was also adopted by [13].

The application in [11] concluded the names of people talking positively versus those who talk negatively about seven topics targeted in news and blogs. The classification was based on a list of words (lexicon) created during the experiment. The application in [14] defines the polarity of emotions in children fairy tales text using a linear classifier. Groups of annotators labeled the stories according to the six basic

ISBN: 978-1-941968-02-4 ©2014 SDIWC

emotions of Ekman. Afterwards, all emotions were grouped under two classes: positive and negative emotions. The classification accuracy reached 69.37%. Different approaches of assigning positive/ negative score to a word out of context were compared in [15]. SVMs and Gaussian Processes were used to test the performance of all the metrics in conjunction. The study suggests a learning approach that combines the various formulae that compute prior posterior polarity. According to [16], for the more difficult multiclass case including a neutral class, accuracy is often below 60% for short messages on Twitter, the social media website. In the later study, the negation of different sentiments was taken into consideration. This practice improved the polarity classification by 5.4% compared to the methods that use bag of words features with NaïveBayes and SVMs.

In the field of emotion detection, the study in [17] considered the identification of the basic emotions in news headlines. Different systems were tried. The first system checks the presence of the WordNet Affect emotion word in the headline and categorizes the text accordingly. The second one was based on the annotation of blogs and the Bayesian classifier. The third system was based on the semantic of the words and allowed the detection of emotion related words. In [18] the author suggested including a semantic feature in the process of sentiment analysis of tweets that enhanced the accuracy of the classification by 6.5%.

The study in [19] targeted the analysis of the tweets containing an emotional hashtag based on the six basic emotions. An emotion labeled tweets corpus has been created (TEC) as well as a word-emotion lexicon. The results suggest that the TEC, after applying a domain adaptation technique, produced better results than the methods used in [17] even when applied in different domains. In addition, Twitter emotional hashtags were used by Gurini et al. [20] to design a recommendation system of friends that share the same sentiment about the topics of interest of each user.

2.2 Sentiment Analysis of Arabic Content

Some studies have been conducted in the field of sentiment analysis of Arabic content. The study in [21] focused on comparing the performance of different classifiers for the Arabic language datasets. Another study [22] compared the performance of the Sequential Minimal Optimization (SMO) and Naïve Bayesian (NB) in the classification task. These algorithms are the suggested classifiers based on excel in their performance; SMO comes first, and then NB. This excel was also highlighted in a polarity classification in [23]. Determining the polarity of opinions in Arabic documents was also the target of [24]. The study proposed a combination between three different classification methods; a lexicon based classifier, followed by the maximum entropy classifier, then K-nearest neighbor classifier. This approach achieved 80% performance accuracy. In addition, a classifier for social networks slang Arabic content was suggested in [25]. This study used 1350 comments as a dataset collected from news channels websites in order to use the SVM classifier to categorize tweets into positive/ negative classes. The training lexicon used was augmented by additional slang terms that enhanced the accuracy by 14%. Another study was concerned with the emotion detection based on the six basic emotions in the Arabic children stories using a computational approach [26]. This study mentions the importance of the word-emotion lexicon and the preprocessing steps that consider the punctuations and the negative words.

The Arabic approaches have some limitations and shortcomings, to our knowledge there is no developed Arabic corpus with instances labeled for emotions. The slang Egyptian dialect has not been investigated yet. Therefore, we took

the approach described in the following section in order to cover some of these limitations.

3 Methodological Approach

The analysis of the data involved several steps: the dataset composition, the annotation process, the data preprocessing, and finally, different classification techniques were tried over the data.

3.1 Data Collection

We collected our data from Twitter, the social media website. The corpus collected includes 1776 tweets from more than 200 users, covering an 18 days period from January 25, 2011 to February 11, 2011. As the usages of words vary from a topic to another [19], we have chosen the Egyptian revolution in 2011 as the topic of concern. It is identified by the hashtag #jan25 on twitter. Tweets ranged from a one word tweet to 140 characters tweet.
We filtered out non Arabic tweets, retweets, tweets including photos or videos. Finally, the corpus was ready for the annotation process.

3.2 Data Annotation

We created surveys where the annotator task was to guess the emotion of the writer based on the provided tweet text. Three annotation runs took place. The 1st run included random Arabic tweets chosen from the collected set. It resulted in 1012 annotated tweets out of 1130 input tweets. The 2nd run included mainly some limited features tweets and it resulted in 609 annotated tweets out of 646 input tweets. The 3rd run was a confirmation run together with an annotation of the tweets according to the communication style used by the tweet writer; aggressive, assertive, or passive style.

The entire tweets collected from Twitter represent an Arabic emotion annotated tweets corpus. The annotators were Egyptians who witnessed the Egyptian revolution in 2011. An average of 15 persons labeled each tweet with the corresponding emotion. The emotions provided are the six most basic emotions [2] [6]. We have excluded the annotated tweets with less than 50% annotators' agreement. Finally, we constructed the Twitter corpus that consisted of 1605 tweets.

3.3 Data Preprocessing

The data preprocessing took place using five different techniques; basic preprocessing, basic preprocessing in addition to the removal of a list of stopwords, Lucene light Arabic stemmer [27], Shereen Khoja Arabic stemmer [28], and modified Khoja Arabic stemmer. The basic preprocessing includes the removal of non Arabic letters, multiple spaces, and punctuation. The list of stopwords is composed of a standard list of standard Arabic words [29]. We have also added to it their equivalents in the Egyptian slang dialect and some additional slang words that appeared in the collected dataset and have no emotion significance. The Lucene light Arabic stemmer eliminates the definite articles and few prefixes and suffixes only. The Khoja stemmer does all the previous functionalities in addition to reducing each word to its root; however, it handles the standard Arabic language only. We modified Khoja stemmer in order to include the Egyptian slang dialect.

3.3.1 Modified Khoja Stemmer

Preprocessing in Arabic language is of great use; especially, each Arabic word can be reduced to its root. The Khoja Stemmer is concerned with affixes removal as well as reducing words to their roots. It includes stemmer text files that contain an Arabic dictionary of roots. Also, lists of stopwords, affixes, as well as a list of strange words are included. After stemming, words should equate one of the roots provided by the stemmer. The Khoja stemmer is designed to support the standard Arabic language only. Our

modifications aimed at customizing the Khoja stemmer to support the Egyptian slang dialect contained in our created corpus. Word segmentation has also been applied; the delimiter used is the space. Other symbols like (.), (,), (;), (:), multiple spaces, and new lines are replaced with one space. The Egyptian slang dialect has some common characteristics that can be emphasized. We handled them the following way:

- **Negation patterns**

In standard Arabic language negated words are preceded by a negation tool or word. For example, the verb hears "YASMA'A" in Arabic is negated if it is preceded by the word no "LA" to be "LA YASMA'A", doesn't hear. However, in slang negation it will be "MAYSMA'ASH". This pattern is formed by the addition of the letter "MEEM" at the beginning of the verb and the letter "SHEEN" at the end of the verb. This pattern has different forms for singular, plural and masculine, feminine person; it is shown in table 1.

Table 1. Negation pattern.

	Subject	Pattern
Present tense	Singular masculine	مبيفعلش
Present tense	Singular feminine	مبتفعلش
Present tense	Plural masculine	مبيفعلوش
Present tense	Plural feminine	مبتفعلش or مبيفعلوش
Past tense	Singular masculine	مفعلش
Past tense	Singular feminine	مفعلتش
Past tense	Plural masculine	مفعلوش
Past tense	Plural feminine	مفعلتش or مفعلوش
Order	Singular masculine	متفعلش
Order	Singular feminine	متفعليش
Order	Plural masculine	متفعلوش
Order	Plural feminine	متفعلوش

We added to Khoja stemmer a Java method that checks the occurrence of such cases. If it is found, the word is reduced to its root instead of being ignored by the original Khoja stemmer. Moreover, it will be preceded by a negation word in the output file to differentiate between the occurrence of word and its negation.

- **Additional suffixes, prefixes, and stopwords.**

We added some slang suffixes and prefixes to Khoja files in order to remove them from the stemmed words through the reduction process. Stopwords are also removed from the text being stemmed. The list of suffixes, prefixes and stopwords is shown in tables **2**, **3**, and **4**.

Table 2. The list of additional suffixes.

Suffixes
لهم، لنا، نكم، ينك
كي، اك
ل، و، ي

Table 3. The list of additional prefixes.

Prefixes
ح، ه، حي، هي
حت، هت
حن، هن

Table 4. The list of additional stopwords.

Stopwords
ليا، لهن، لهم، له، لكي، لك
دول، ديه، دي، دا، دة، ده
دلوقتي، دلوقت، الآن، الان
انتي، باني
وكنتوا، كنتوا، وكنتو، كنتو، وكنتم، كنتم، وكانوا، كانوا، وكنت، كنت
بوك، فيس، تويتر، اللي، شوية، بس

- **Additional definite articles.**

For abbreviation, some letters are sometimes added before the definite article "AL". For example, the letter "AIN" to mean on top of, or

the letter "SEEN" to mean the night of. Hence, a new form of the article appears. We added these forms to the definite articles text file in the stemmer.

- **Slang representation of periodontal letters.**

The Arabic language includes three periodontal letters, where two of them are replaced by other letters in the Egyptian slang language. The letter "THEH" is replaced by the letter "TEH", and the letter "THAL" is replaced by the letter "DAL". We added a Java method to check this case in order to reach the correct root of the word.

Moreover, three semantic properties were supported:

- **Word pairs**

In Arabic language some word pairs like "ALHAMD LELLAH" which means thanks God and "ALLAH AKBAR" which means God is the greatest, are used to express gratitude to God and happiness. Mistakenly, sometimes people write them as a single word. To avoid stemmer misinterpretation of such word pairs, we added a method to the code that checks the presence of these pairs and replaces them with same word in the output file.

- **Negation form**

As mentioned in the negation pattern property described earlier, any word preceded by a negation tool or is in the slang negation pattern form should not be replaced by its root. Instead, it should be replaced by the negation of this word in the output file. We added a negation words list (standard and slang), so that it could be checked during the reduction process. If the algorithm finds a word of this list in front of another word, both will be replaced by the root of the word preceded by a specific negation word.

Table 5. The list of negation words that are replaced with the negation word (لا) in the output file:

Negation words
بلا، فلا، ولا، لا
فلم، ولم، لم
فلن، ولن، لن
فليس، وليس، ليس
منغير، بغير، بدون

- **Disapprobation words**

In the Egyptian slang language, some words are used to express the same meaning. For example, "HOWA EZAY", "HOWA EIH" both denotes disapprobation. The form of these words is usually a word denoting a question followed or preceded by a stopword, or a negation word. We added a method in the code to check for such words and replace them with the same word in the output file.

These additions were adopted in order to facilitate the classification job and enhance its performance. Using the modified Khoja stemmer as the preprocessing technique for the first annotation run tweets (1012 tweets), we extracted some emotion related attributes in order to form a seed list for a sample word-emotion lexicon.

3.4 Sample Word-Emotion Lexicon

Weka selects attributes option was used in order to extract the most effective features included in the first 1012 annotated tweets. The feature selection algorithm used was **BestFirst**. The extracted features formed the base of a word-emotion lexicon. In addition, the lexicon has been extended with some manually crafted emotion related words that were manually extracted from these tweets. The direct annotation of words performed in this way usually performs better than other methods according to [19]. The created lexicon was used in the collection of the second annotation run tweets. This sample lexicon was used in order to create a limited features environment for

further experiments as will be shown in section 4.

3.5 Data Classification

Finally, different data classification techniques were tried out. Weka software has been used for the SMO classifier, which is a simplification of the SVM classifier, and for the NaïveBayes classifier. And, a simple search and frequency algorithm based on the extracted sample word-emotion lexicon was also created. This algorithm counts the number of each emotion related words in the tweet. Then, it decides the emotion category of the tweet based on the emotion receiving the highest count. Ten folds cross validation was applied for the learning based algorithms. A comparison between the performances of the different classification algorithms has been held. Further, we conducted a comparison between the effects of the five different preprocessing techniques. Two different environments have been tried the random tweets environment and the limited features tweets environment.

4 Experiments and Results

The evaluation of the performance involves the calculation of the precision (P) and recall (R). They are calculated as follows:

$$P = \#\text{correct guesses} / \#\text{total guesses} \quad (1)$$

$$R = \#\text{correct guesses} / \#\text{total} \quad (2)$$

The number of correct guesses is the number of tweets marked correctly as expressing an emotion X by the classifier. The total guesses, is the total number of tweets that are marked by the classifier as expressing the emotion X (including correct and wrong guesses). The total number is the number of tweets expressing the emotion X in the dataset. Moreover, F is the balanced F-score which is calculated using the following formula (3).

$$F = 2PR / (P+R) \quad (3)$$

4.1 Random Tweets Tests

The first 1012 tweets that were randomly chosen were subject of different tests. They consist of 259 anger tweets, 127 disgust tweets, 149 fear tweets, 271 happiness tweets, 110 sadness tweets, and 96 surprise tweets. The first test was the comparison between the classification performance of the SMO and NaïveBayes classifiers.

4.1.1 SMO vs NaïveBayes

The first 1012 tweets were stemmed using the modified Khoja stemmer. Cross validation test with 10 folds run over these tweets using both classifiers. The weighted average results of both classification algorithms are shown in table 6. Comparing the overall performance of both classification algorithms, the experiment demonstrates that the SMO classifier outperforms the NaïveBayes classifier by almost 5%. Therefore, in the rest of the tests, the SMO classifier was chosen.

Table 6. The weighted average results of SMO vs. NaïveBayes performance.

Classifier	P	R	F
NaïveBayes	0.399	0.391	0.394
SMO	0.442	0.451	0.441

4.1.2 The Five Preprocessing Techniques

The five preprocessing techniques have been compared. The weighted average classification results using the basic preprocessing (BP), basic preprocessing and removal of stopwords (BPRS), Lucene light Arabic stemmer (LLAS), Khoja Arabic stemmer (KAS), and modified Khoja Arabic stemmer (MKAS) are represented in table 7.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Table 7. The weighted average classification results of the first 1012 random tweets with different preprocessing techniques.

Preprocessing Technique	#attributes	P	R	F
BP	6007	0.423	0.418	0.402
BPRS	5875	0.41	0.409	0.387
LLAS	5371	0.423	0.418	0.402
KAS	1283	0.394	0.409	0.397
MKAS	1450	0.442	0.451	0.441

BP technique resulted in 6007 attributes. The classifier took advantage of the variety of attributes and related a large number of features to each emotion category. That was done even with the words that do not really have emotion significance, for example, the words contained in the stopwords list. BPRS technique resulted in 5875 attributes. The attributes that were erroneously considered related to emotions in BP technique, were unique for each corresponding emotion class. When they were removed in BPRS technique, fewer words were associated to each emotion class, hence, a decrease in the overall performance of the classifier.

LLAS technique only removes the definite articles and few prefixes and suffixes. Therefore, the number of attributes decreased to 5371. However, this reduction did not enhance the overall classification results compared to the results of BP technique. The KAS technique reduced the number of attributes to 1283. This reduction is due to the stopwords list eliminated, in addition to the reduction of each word to its root. In this manner, after stemming, many words are represented with the same root, hence, the same attribute. Moreover, this technique removes any non stemmed word even if it is not from the stopwords list except the list of the strange words, which has been defined in the stemmer files before the test. Therefore, the classifier did not depend on relating many attributes to the emotion category. Instead, it depended on less attributes that were more frequently repeated in the different tweets. The MKAS that we developed takes into consideration the Egyptian slang language words that occurred in the tweets as mentioned in section **3**. Thus, the number of attributes that resulted from the implementation of this technique is 1450 attributes. This increase, compared to KAS technique can be related to the inclusion of some Arabic slang words that were removed by the KAS technique. Compared to BP technique, our MKAS technique enhanced the classification results by 3.9%. While, compared to the original KAS, it resulted in 4.4% enhancement in the overall classification results.

The attributes specified in the preprocessing stage are the features that the classification mainly depends on. We have chosen from them the most significant attributes as mentioned in **3** to be the seeds of a sample Arabic word-emotion lexicon.

4.2 Limited Features Tweets Tests

The random choice of tweets compared to the size of the sample set collected might not show the classification performance neatly. Therefore, the tweets that include the extracted features (the sample word-emotion lexicon) were selected to form a set of limited features tweets. The assumption is that allowing more repetition of words and separability of features related to each class would compensate for a larger dataset. Thus, the concept can be generalized for bigger size sets. Tweets containing words from the sample word-emotion lexicon were grouped together. A total of 1000 limited features tweets collected from the total 1605 tweets set that were finally available. They have been subject of the same tests as the previously selected random tweets.

4.2.1 The five preprocessing techniques

We checked the effect of the different preprocessing techniques on the classification performance for the limited features tweets as

shown in table **8**. Taking the BP technique results as a reference, the usage BPRS technique enhanced the results by 1.1%. LLAS technique resulted in 1.7% enhancement. KAS technique increased the overall performance by 14.6%. Moreover, it has been shown that the MKAS enhanced the results by 17% compared to the classification using BP technique. Thus, our modified Khoja stemmer still outperforms the other preprocessing methods.

Table 8. The weighted average classification results of the limited features tweets using the five different preprocessing techniques.

Preprocessing Technique	Attributes	P	R	F
BP	5952	0.555	0.553	0.539
BPRS	5833	0.571	0.562	0.55
LLAS	5331	0.568	0.566	0.556
KAS	1277	0.691	0.685	0.685
MKAS	1433	0.716	0.71	0.709

Finally, the total dataset was classified using the SMO classifier. Moreover, we checked the performance of the simple search and frequency algorithm based on the sample word-emotion lexicon.

4.3 Total tweets set test

The total tweets set consists of 409 anger tweets, 204 disgust tweets, 285 fear tweets, 340 happiness tweets, 201 sadness tweets, and 166 surprise tweets. It was classified using the SMO classifier with cross validation option. It resulted in a weighted average balanced f-measure of 0.531, weighted average precision of 0.535, and weighted average recall of 0.535.

The SMO classifier with train/test option was investigated. The 1012 tweets of the first run were given to the algorithm as training set. In addition, the second run tweets (609 tweets) were entered as testing data. The MKAS preprocessing technique was used. The results of using the sample word-emotion lexicon search and frequency (SF) algorithm in the classification of the same 609 tweets were compared to the SMO classifier results. As shown in table **9** the experiment demonstrates that the results of the algorithm that uses the sample word-emotion lexicon exceed the results of the trained SMO classifier.

Table 9. The weighted average classification results of SMO vs. SF algorithm.

Emotion	Algorithm	P	R	F
Anger	SMO	0.407	0.641	0.498
	SF	0.532	0.885	0.665
Disgust	SMO	0.333	0.253	0.287
	SF	1	0.462	0.632
Fear	SMO	0.648	0.431	0.504
	SF	0.838	0.65	0.732
Happiness	SMO	0.422	0.713	0.53
	SF	0.542	0.888	0.673
Sadness	SMO	0.431	0.272	0.334
	SF	0.731	0.533	0.616
Surprise	SMO	0.48	0.333	0.393
	SF	0.816	0.431	0.564

After that, we checked our assumption that there exists a correlation between the emotion expressed in the sentence and the communication style used by the writer.

4.3.1 Communication style in emotion detection

According to the communication style analysis of the total 1605 tweets shown in table **10**, we can use the exclusion of some emotions upon the presence of a specific communication style.

Table 10. Communication style in 1605 tweets.

	Aggressive	Assertive	Passive
Anger	76.53%	16.91%	6.85%
Disgust	83.33%	9.31%	7.35%
Fear	3.4%	38.11%	58.39%
Happiness	3.83%	86.73%	9.44%
Sadness	32.18%	40.1%	27.72%
Surprise	37.95%	36.75%	25.3%

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Looking at the minorities, anger, disgust, and happiness are rarely expressed using the passive communication style. Moreover, few fear and happiness tweets are expressed in the aggressive communication style. In addition, only 9.31% tweets of the disgust emotion category are expressed using the assertive communication style.

The results in table **11** show the issue from communication style point of view.

Table 11. Communication style percentage distribution excluding surprise and sadness.

	Anger	Disg-ust	Fear	Happin-ess
Aggressi-ve	61.86%	33.6%	2.0%	2.5%
Assertive	13.88%	3.88%	26.65%	60%
Passive	11.57%	6.2%	69.0%	13.22%

If we adopt [30] emotional model that suggest that disgust is a secondary emotion of anger, then 95.46% of the aggressive communication style tweets will be expressing anger. Although the assertive communication style is used in the expression of several emotions, it is more associated with the happiness emotion category.

5 Discussion

Our created corpus size is comparable to the OCA corpus size in [23] which consists of 500 reviews, divided into positive and negative ones, in addition to the 1143 corpus in [24] and the 1000 tweets datasets in [18]. Classification algorithms like SMO and NaïveBayes that proved good performance for English text categorization are applicable for Arabic text as well. Similarly to the English language case our experiment demonstrates that SMO classifier outperforms the NaïveBayes classifier for Arabic text categorization by 5.4%. These test results support the claim in the text polarity classification in [23] and [22]. Moreover, the same finding was valid for English language polarity test in [14] that used a linear classifier and in [31]. In the text categorization, the support vectors idea is more effective as it separates the classes with the largest margin and does not depend on the individual words probabilities like the NaïveBayes classifier.

We compared the classification results of our set of tweets using the modified Khoja stemmer and the SMO classifier with the classification results of the 1000 English news headlines in [19]. The results of the comparison concerning the precision and recall results of the different emotions are shown in table **12**. In this table, we refer to the study [19] by the number 1 and to our study with the number 2.

Table 12. P. & R. comparison between our study and [19].

	Study	P	R
Anger	1	0.493	0.265
	2	0.474	0.597
Disgust	1	0.421	0.186
	2	0.488	0.389
Fear	1	0.635	0.437
	2	0.582	0.533
Happiness	1	0.54	0.367
	2	0.623	0.677
Sadness	1	0.525	0.367
	2	0.479	0.345
Surprise	1	0.443	0.292
	2	0.545	0.506

Although the test sets are not similar, and the language used is different, the previous comparison gives a rough estimate of how good the classification performance is.

Moreover, we have added some code that adds our additional features to Khoja stemmer in order to support the Arabic Egyptian dialect and enhance the classification results. It has been shown that our modifications improved the overall classification results by almost 4.4% compared to the original khoja stemmer. Our modified Khoja stemmer solved some of the problems mentioned by Al-Khalifa in [32] such as the negation forms, the extra white spaces,

and the spelling mistakes found in the tweets. The preprocessing steps help the classifier to narrow down its feature scope and avoid noise features.

The experiment demonstrates that the preprocessing was more effective in case of limited features where the likelihood of word repetition is higher. Therefore, the classifier can associate words with emotion categories in the training phase and find them frequently in the test phase. Thus, if the training set increases to include tweets that enclose all the emotion related words of a certain language, the preprocessing is supposed to be more effective.

It has been shown that the sample word-emotion lexicon created based on the significant attributes in the tweets enhanced the overall performance of the classification by 22.27% compared to the SMO classification using the train/test option. Comparing this result with [19] and [17], we can support their claim that a word-emotion lexicon would easy the emotion detection task. In [17] the usage of emotion words has positively affected the overall classification performance by 4.35% compared to the blogs training classification performance. Moreover, the emotion related words extracted from microblogs in [19] enhanced the classification results by 19.05 % compared to the original classification.

Yet, the experiment demonstrates that the usage of the SMO classifier and the word-emotion lexicon is a tradeoff. In case of the availability of a dataset with a variety of words, the sample word-emotion lexicon will not be as effective. That is due to the fact that the SMO algorithm uses almost all contained words as features, and it creates the classification model accordingly. This big number of words enriches the model and gives it the ability to deal with random tweets. On the other hand, it has been shown that the usage of a sample word-emotion lexicon has a great benefit if the test data contains these words. However, if such this sample word-emotion lexicon is used to classify a random test set, it will not perform as good as the SMO classifier. Therefore, in order to have the benefit of the lexicon in all cases, a full word-emotion lexicon should be developed and used.

The experiment demonstrates that a reciprocal relationship exists between the emotion and the communication style. It has been shown in section 4.3.1 that a correlation exists between the emotion expressed and the communication style. Based on the analysis of the data, we can exclude the fear and happiness emotions if the communication style is aggressive. Further, we can exclude the disgust emotion if the communication style is assertive or passive. More likelihood should be given to the fear emotion in case of passive communication style. More likelihood should be given to the happiness emotion in case of assertive communication style.

6 Limitations

While analyzing the results of this study, certain limitations need to be taken into account. The dataset that this study focused on is the Egyptian revolution in 2011 topic. Moreover, the size of the dataset (1605 tweets) was not big. However, the annotation of each tweet is very accurate due to two factors. The first one is the doubled number of annotators compared to other studies like [14] and [17]. The second factor is that each tweet underwent more than one annotation run. Moreover, the vocabulary used to express emotions may vary from topic to another according to [19]. Thus, the performance of classification models represented in this study for other topics is not guaranteed. Finally, the study focused on the Arabic language tweets. The preprocessing techniques are customized for the Arabic Language words (standard and slang Egyptian dialect).

ISBN: 978-1-941968-02-4 ©2014 SDIWC

7 Conclusion

This study tackled the automatic detection of emotions in textual parts of Twitter the social media website. This process has applications for business development, user interface design, content creation, emergency response, among others. The study was concerned with the capability of classifying Arabic text considering the standard and the slang Egyptian dialect. The dataset under consideration is composed of Arabic text messages collected from Twitter during January/February 2011. The focus topic is labeled with #Jan25 referring to the Egyptian revolution in 2011. A corpus of emotion annotated tweets was developed as well as a sample word-emotion lexicon. Five different preprocessing techniques have been compared. New features were added to Khoja stemmer in order to support some of the slang Egyptian dialect. It has been shown that the use of this modified Khoja stemmer has been associated to the best classification performance. Moreover, the experiment demonstrates that our simple search and frequency algorithm performed better than the SMO classifier in the limited features environment.

We are contributing to the research area of the emotion detection of Arabic content by showing that emotions can be automatically detected from tweets after performing Arabic language related language preprocessing steps. The experiment demonstrates that the preprocessing steps added to Khoja stemmer improved the classification results by 4.4% compared to the original Khoja stemmer performance. In addition, it has been shown that our sample word-emotion lexicon enhances the emotion detection results by 22.27% compared to the SMO classification using the train/test option. Finally, it has been shown that the communication style is closely related to the emotion expressed in case of anger, disgust, fear, and happiness categories. The relationship can be thought of as a reciprocal one.

8 Future Work

The results of this study point to different interesting directions for future work. First, a complete Arabic word-emotion lexicon can be developed. That could be done by considering tweets from different topics and annotating them by emotion. Or, it could be done by going through all the words in the Arabic dictionary and annotating those that have emotion significance.
The WordNet affect, the NRC, and the lexicon extracted from the TEC can be taken as a reference in order to check the inclusion of all corresponding Arabic emotion related words. Moreover, the development of a stemmer that supports the different Arabic slang dialects would be of a great effect. In addition, the expansion of the Arabic tweets corpus with more instances labeled for emotions should be considered. This step would help the training task and make it better for the SMO classifier. Finally, the development of an automatic system that detects the communication style from Arabic text based on the structure of the sentence. This system could facilitate the emotion detection task.

7 REFERENCES

[1] M. S. Thambirajah. 2005. "Psychological Basis of Psychiatry." Churchill Livingstone.

[2] P. Ekman. 1992. "Are there basic emotions. Psychological Review." Psychological Review, volume 99, pages 550–553.

[3] J. A. Russel. 1994. "Is there universal recognition of emotion from facial expression? a review of the cross-cultural studies." Psychological Bulletin, volume 115, pages 102–141.

[4] P. Ekman. 2000. "Basic emotions." In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion, pages 45–60. Sussex, U.K.: John Wiley and Sons, Ltd.

[5] Arabic Web Days. 2012. http://www.arabicwebdays.com/. Online; accessed May-2013.

[6] P. Ekman. 1971. "Universals and cultural differences in facial expression of emotions." Lincoln: University of Nebraska Press, volume 19, pages 207–282.

[7] B. Pang and L. Lee. 2008. "Opinion mining and sentiment analysis." Found. Trends Inf. Retr., volume 2, pages 1–135.

[8] V. Hatzivassiloglou and K. McKeown. 1997. "Predicting the semantic orientation of adjectives." In Proceedings of the Joint ACL/EACL Conference, pages 174–181.

[9] J. Wiebe. 2000. "Learning subjective adjectives from corpora." In Proceedings of the Seventeenth National Conference on Innovative Applications of Artificial Intelligence, pages 735-740. AAAI Press.

[10] A. Valitutti, C. Strapparava, and O. Stock. 2004. "Developing affective lexical resources." PsychNology, volume 2, pages 61-83.

[11] N. Godbole, M. Srinivasaiah, and S. Skiena. 2007. "Large-scale sentiment analysis for news and blogs." In Proceedings of the International Conference on Weblogs and Social Media (ICWSM).

[12] B. Pang, L. Lee, and S. Vaithyanathan. 2002. "Thumbs up?: Sentiment classification using machine learning techniques." In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP '02), volume 10, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

[13] T. Nasukawa and J. Yi. 2003. "Sentiment analysis: capturing favorability using natural language processing." In Proceedings of the 2nd international conference on Knowledge capture (K-CAP '03), pages 70–77, New York, NY, USA. ACM.

[14] C. O. Alm. 2005. "Emotions from text: Machine learning for text-based emotion prediction." In Proceedings of HLT/EMNLP, pages 347–354.

[15] M. Guerini, L. Gatti, and M. Turchi. 2013. "Sentiment analysis: How to derive prior polarities from sentiwordnet." In EMNLP, pages 1259–1269. ACL.

[16] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. 2013. "Recursive deep models for semantic compositionality over a sentiment treebank." In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631-1642, Seattle, Washington, USA. Association for Computational Linguistics.

[17] C. Strapparava and R. Mihalcea. 2008. "Learning to identify emotions in text." In Proceedings of the 2008 ACM Symposium on Applied Computing (SAC'08), pages, 1556-1560, New York, NY, USA. ACM.

[18] H. Saif, Y. He, and H. Alani. 2012. "Semantic sentiment analysis of twitter." In International Semantic Web Conference (1), pages 508–524.

[19] S. Mohammad. 2012. "Emotional tweets." 2012. The First Joint Conference on Lexical and Computational Semantics, Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 246–255. Association for Computational Linguistics.

[20] D. F. Gurini, F. Gasparetti, A. Micarelli, and G. Sansonetti. 2013. "A sentiment-based approach to twitter user recommendation." In RSWeb@RecSys.

[21] M. Al-diabat. 2012. "Arabic text categorization using classification rule mining." In Applied Mathematical Sciences, volume 6, pages 4033-4046.

[22] B. Al-Shargabi,W. Al-Romimah, and F. Olayah. 2011. "A comparative study for arabic text classification algorithms based on stop words elimination." In Proceedings of the 2011 International Conference on Intelligent Semantic Web Services and Applications (ISWSA '11), pages 1–5, New York, NY, USA. ACM.

[23] M. Rushdi-Saleh, M. T. Martn-Valdivia, L. A. Lpez, and J. M. Ortega. 2011. "Bilingual experiments with an arabic-english corpus for opinion mining." In Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, and Nicolas Nicolov, editors, RANLP, pages 740–745. RANLP 2011 Organising Committee.

[24] A. El-Halees. 2011. "Arabic opinion mining using combined classification approach." International Arab Conference on Information Technology (ACIT'2011), volume26, Riyadh, Saudi Arabia.

[25] T. H. Soliman, A. Hedar, M. Ali, and M. M. Doss. 2013. "Mining social networks' arabic slang comments." In Proceedings of IADIS European Conference on Data Mining 2013 (ECDM'13).

[26] A. F. El-Gohary, T. I. Sultan, M. A. Hana, M. M. El-Dosoky. 2013. "A Computational Approach for Analyzing and Detecting Emotions in Arabic Text." In International Journal of Engineering Research and Applications (IJERA), volume 3, pages 100-107.

[27] Apache Software Foundation. 2010. http://lucene.apache.org/core/old_versioned_docs/versions/3_0_3/api/contrib.-analyzers/org/apache/lucene/analysis/ar/ArabicAnalyzer.html. Online; accessed June-2013.

[28] S. Khoja. 2001. http://zeus.cs.pacificu.edu/shereen/research.htm. Online; accessed June-2013.

[29] Arab geeks. 2009. Arabic stop words. http://arabicstopwords.sourceforge.net/. Online; accessed April-2013.

[30] W. G. Parrott. 2001. "Emotions in Social Psychology: Essential Readings." Key readings in social psychology. Psychology Press.

[31] T. Joachims. 1998. "Text categorization with support vector machines: Learning with many relevant features." In European Conference on Machine Learning (ECML), pages 137-142, Berlin. Springer.

[32] L. Albraheem and H. S. Al-Khalifa. 2012. "Exploring the problems of sentiment analysis in informal arabic." In Proceedings of the 14th International Conference on Information Integration and Web-based Applications (IIWAS'12), pages 415–418, New York, NY, USA. ACM.

Towards Applying Support Vector Machine Algorithm in Employee Achievement Classification

Hamidah Jantan, Norazmah Mat Yusoff and Mohamad Rozuan Noh
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM) Terengganu,
23000 Dungun, Terengganu, Malaysia
{hamidahjtn,norazmah}@tganu.uitm.edu.my, rozuannoh@yahoo.com

ABSTRACT

Human capital is the key factor to maintain the competitiveness of an organization by having enough right people with the right skills. In technology advancement, machine learning technique can be used in order to identify the right employee for the right task by classifying their performance achievement. Support Vector Machine (SVM) is a powerful supervised machine learning technique for classification because it uses kernel trick with the ability to build expert knowledge for the problem via kernel engineering process. In this study, Sequential Minimal Optimization (SMO) algorithm from SVM technique is the chosen method due to its capability to solve most of convex optimization problem. This study consists of four phases; data collection, data preparation, model development and model evaluation. In the experimental phase, selected academician performance achievement data in Malaysian Higher Institution have been used as the training dataset based on 10-fold cross validation. Several experiments were carried out by using different set of training and testing datasets to evaluate the accuracy of the model. As a result, the accuracy of the proposed model is considered acceptable and needs further enhancement. For future work, to enhance the accuracy of the proposed model, a comparative study should be conducted using other SVM algorithms such as Grid Search and Gabriel graph algorithms that focus on reducing the size of a training set.

KEYWORDS

Employee's Achievement, Categorization, Support Vector Machine (SVM), Sequential Minimal Optimization (SMO).

1 INTRODUCTION

Globalization and fast technological advancement have changed the survival environment of an organization. Human capital within an organization is the key factor to maintain the competitiveness of the organization. In any organization, its employees should be deployed in the appropriate locations at the appropriate point of time via talent management activities [1-3]. This process involves a lot of managerial decisions that counter through employees' performance achievement analysis which is sometimes very difficult, uncertain and challenging. It also depends on various aspects related to their profession criteria such as academic background, experiences, quantity and quality of work, knowledge, skill, contributions and etc. Therefore, the most challenging endless task for Human Resource (HR) professionals is managing their employees [4].

Nowadays, a machine learning technique has given a great deal of concern and attention in information industry because of its ability to produce intelligent decision that implements Knowledge Discovery in Database (KDD) approach. This is due to the wide accessibility of enormous amount of data and the important need to turning such data into useful information as intelligent knowledge [5]. This approach has been applied in many areas such as manufacturing, engineering, medical, finance, marketing, health care, customer

relationship and etc. However, this approach dose not really attract researchers in HR and HR decision support application that use this approach is also quite rare [6, 7]. HR databases have rich of hidden and valuable knowledge to help decision making in talent management especially for employee achivement classification [7-9].

In technology advancement, there are many supervised machine learning classifications that can be used for classification such as Decision Tree, Artificial Neural Network (ANN), Random Forest, Naïve Bayesian, RBF Network, Artificial Immune System (AIS), Support Vector Machine (SVM) and etc. This study endeavours to investigate the effectiveness of SVM technique in identifying the required data pattern for employee achievement classification. SVM is considered as a powerful technique in classification and leads to increase the performance in pattern recognition, regression, estimation and etc.[10]. Besides that, SVM is known as the most robust and accurate method among the well-known algorithms such as back-propagation neural network (BPN), k-means and C4.5 algorithms [11, 12]. SVM can be used for classification with optimization ability for complex non-linear decision boundaries. In this study, Sequential Minimal Optimization (SOM) is the chosen SVM algorithm because it is known as an efficient classifier in solving optimization problem. SOM can be considered as the state of the art approach in non-linear SVM [13]. Due to these reasons, this study aims to suggest SVM classification model using SOM algorithm for academic performance achievement via several experiments on selected datasets. The rest of this paper is organized as follows; the second section discusses on the related work and studies of motivation in regards to the employee performance achievement, the classification in Machine learning, SVM techniques and application , and discussion on SMO algorithm. The third section describes the experiment that was conducted in this study. Then, the fourth section discusses the results and discussions. Finally, the paper ends with section five where the concluding remarks and future research directions are identified.

2 RELATED WORKS

2.1 Performance Achievement

Performance of employee is considered as an individual natural proficiency, skill, ability or aptitude that can be seen as the result of activities over a given period of time. The achievement of performance is a crucial issue in talent management to enable the success of an organization. This achievement becomes the most important resources of regional economy and social development. Besides that, it also provides important source of information and decision making basis and plays significant roles in future direction of an organization [14, 15]. In addition, performance achievement evaluation is considered as a mechanism to determine the requirements needed for high competence employee, position level and profession track. Besides that, this process can understand the gap between basic requirements and demand of employees in various industries. This task can also help Human Resource professionals to determine the training needs, professional setting and structure of human capital development in an organization.

The process to determine the right employee for the right position means the process of matching his knowledge, skills and personality with the job requirements, which is known as talent management [16]. Therefore, the process of classifying employee performance achievement can be considered as an alternative method to help HR professionals in matching their employees to the right task or position more effectively. Performance achievement is classified based on various factors and criteria such as quantity and quality of task, skill, knowledge, attitude and etc. Nowadays,

ISBN: 978-1-941968-02-4 ©2014 SDIWC

advancement of computing technology, performance classification in HRM field can be implemented more effectively through knowledge discovery in database approach. Databases in Human Resource Management (HRM) department can provide rich resources for knowledge discovery and decision support tools. Evaluation of employee which has high performances on capability, knowledge, skill, and other abilities plays significant roles in the success of an organization [15].

Talent management is an important issue in any organization where any supported activities for this task will directly influence future direction for successful organization [17]. Nevertheless, competitions among organizations also become the competitions among employees. The competition can be measured through their performance achievement by determining the best candidate to fill up the job opportunities. Process to find the high performance achievement candidate means to find the most matching candidate with his knowledge, skills and personality to job requirements. Employee performance achievement classification involved a process of finding the right talent besides enhancing the most suitable employee's skill and knowledge. There are several studies on employee performance classification as a part of performance achievement evaluation activities [18, 19].

Recently, statistical and soft computing technique was considered as an efficient method used in performance classification due to the ability in producing more accurate results [20]. As an example, research that used statistical technique for project construction performance evaluation using Data Envelopment Analysis (DEA) and AHP proposed an improved mathematical model of DEA for quantitative analysis [21]. Besides that, classification using soft computing technique such as for job performance evaluation on ability, attitude and team spirit in commercial banks was conducted using Fuzzy logic technique [22]. In advancement of classification approaches, there are some studies on machine learning technique for talent classification which has been performed to enhance the ability of the previous proposed method in classification [16, 23, 24].

2.2 Classification in Machine Learning

Databases are rich with hidden knowledge that can be used in decision making process in order to produce intelligent decision. Intelligent decision refers to the ability to make decision which is quite similar to human decision. Clustering, classification and prediction in machine learning are known as common methods to support intelligent decision making. In addition, classification in machine learning known as supervised learning method that can be used to extract models describing important data classes or to predict future data trends [5]. In classification, there are two phases involved; the first phase is learning process where the training data are analyzed by the selected classification algorithm. The learned model or known as classifier is presented in the form of classification rules or patterns. The second phase is model evaluation to estimate the accuracy of the proposed model. If the accuracy is considered acceptable, the rules can be applied to unseen data or untrained data.

Nowadays, there are many classification methods proposed by researchers in machine learning, pattern recognition, and statistics [25]. Some of the techniques that are being used for classification in Machine learning are decision tree, Bayesian methods, Bayesian networks, rule-based algorithms, neural network, support vector machine, association rule mining, k-nearest-neighbor, case-based reasoning, genetic algorithms, rough sets and fuzzy logic [26]. Support Vector Machine (SVM) is the most powerful supervised machine learning technique which has a simple structure and good classification ability [10]. Besides that, SVM is also known as the suitable

algorithm in machine learning for classification task especially on both linear and non-linear decision margins.

2.3 SVM Techniques and Applications

SVM technique was introduced by Cortes and Vapnik in 1995 with many advantages such as it can perform well for data sets that have many attributes resolving the small sample, non-linear and high dimensional pattern recognition [27, 28]. Moreover, SVM has no upper limit on the number of attributes and uses kernel trick which the model can be built within expert knowledge on the problem by adjusting the kernel. Besides that, SVM is the most competent methods for training which can produces high accuracy of model. In two-class learning, the aim is to find the most suitable classification function in differentiating between the members of two categories in a training dataset. Therefore, in a linearly separable dataset, a linear classification function will corresponds to the separating hyperplane that passes through the middle of the two categories by separating it into two different categories. Once this function is identified, new data can be classified by simple testing task in assigning the data to the categories that they belong to. [29]. SVM implements the idea that vector is nonlinearly mapped to a very high dimension future space. In this feature space, a linear separation surface is created to separate the training data by minimizing the margin between the vectors of the two classes [30].

Due to many such linear hyperplanes, SVM technique has found the most suitable function by maximizing the margin among the two categories. Intuitively, the margin is show the amount of space, or separation among the two categories as defined by the hyperplane. The margin corresponds to the minimal distance between the closest data points to a point on the hyperplane. SVM insists to identify the maximum margin hyperplane because it offers the best generalization ability. This will indicates not only the best classification performance or correctness on the training data, but also much room for the correct classification of the future data [11]. SVM algorithm for training used kernel parameter to separate data between the hyperplane which is the input data are called input space and the output data in hyperplane are called feature space. There are four types of kernel in SVM: linear kernel function; polynomial kernel function; Gaussian (RBF) and S-type (sigmoid kernel function). After data separating process is done, the minimal margin between vector and hyperplane is calculated.

SVM technique for classification has been applied in many areas such as in industrial nuclear energy, tourism, image processing and multimedia, geologic and medical [13, 28-32]. The most common field uses SVM algorithm is in medical fields such as to classify Alzheimer's disease from whole-brain anatomical MRI [30]. Alzheimer's disease patients would consequently gain from early and accurate diagnosis of Alzheimer's disease. The other application of SVM is handwritten chemical symbol classification [29]. However, the use of SVM algorithm in HRM is quite rare. There are some studies on the use of SVM technique in HRM such as talent classification using radial basic function for employee recruitment [27]; managing talent using SVM algorithm and Class-Attribute Contingency Coefficient (CACC) to enhance the traditional SVM approach [10]; forecasting regional talent demand using SVM and Principal Component Analysis (PCA) to improve talent demand prediction [14]; and employee selection using Affinity Propagation and SVM Sensitivity Analysis [33]. In addition, machine learning for knowledge discovery is a domain driven that is based on problem to solve using various techniques. Due to that reason, this study attempts to solve employee performance achievement classification for promotion purposes using SMO algorithm as an alternative

approach to help HR professional in managing academic talent in higher institution.

2.4 Sequential Minimal Optimization (SMO)

SMO is an algorithm for solving efficiently the optimization problem arises during the process of SVM training. SMO can be considered as the state–of–the–art approach in a non-linear SVM [13, 34]. The kernel function is adopted in order to implement SVM-based classifiers in mapping into a higher dimensional space. A soft-margin SVM is trained by solving a quadratic programming problem. SVM will train all training datasets using SMO algorithm to generate classification model. SVM step-by-step process is explained below and shown in Fig.1:

a) Separate data from input space (input data) to feature space (output data) by using kernel function or kernel parameter.
b) SVM model is trained in solving a quadratic programming problem using SMO algorithm.
c) The minimal margin between two vectors is calculated using SMO algorithm. This margin will be used in training the whole training datasets to generate classification model.
d) Test the model by using testing data (untrained data) to get the accuracy of the proposed model.

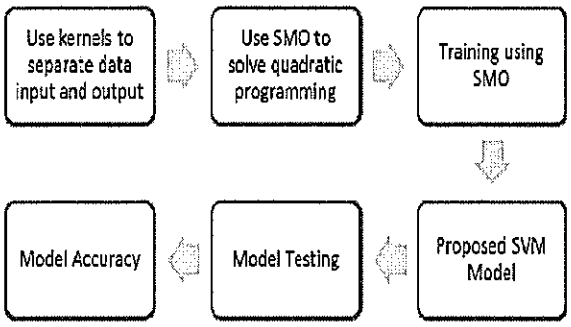


Figure 1. SOM Algorithm in SVM Classification Technique

3 EXPERIMENT SETUP

SVM classification process has two phases; the first phase is learning process where training data are analyzed using selected SVM algorithms. The second phase is classification process; where testing datasets are used to estimate the accuracy of the proposed classification model. If the accuracy is acceptable, the rules can be applied to new data (untrained data) for classification.

In the experimental phase, this study is aimed to discover employee's achievement performance patterns in academics performance databases using SOM algorithm in SVM technique. There are three phases involved i.e. talent model data preparation, model development and model analysis, as shown in Fig. 2.

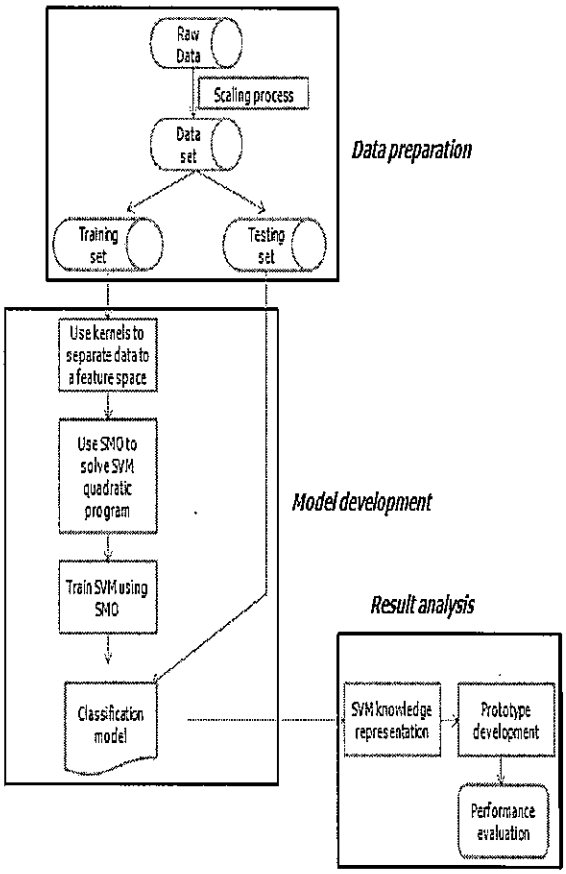


Figure 2. SVM Classification Model Development

The process of classification uses input variables from performance achievement criteria; and the outcome is the employee's performance pattern that shows the status of promotion (Table 1).

TABLE 1. ACADEMIC DATA DESCRIPTION

Criteria	Attribute Name
Demographic (5)	Academic Position
	Length of Service (LS)
	LS before Senior Lecturer
	LS before Associate Professor
	LS before Professor
Research (2)	Number of Research
	Research leader
Publication and Conference (4)	National Journal
	International Journal
	National Conference
	International Conference
Administrative (1)	Administrative post
Contribution (2)	Research Award
	Service Award

In this study, dataset used contained 14 attributes for related performance criteria (Table 1) from 256 academics performance achievement data. SVM technique is a supervised learning method where the target class is known. Due to that reason, dataset used in this experiment contained 13 attributes for training and testing and one attribute for class. Academic position (professor, associate professor, senior lecturer and young lecturer) was selected as a target class that consisted of four levels for classification. The clean data was divided into training and testing dataset. The training dataset will use SVM algorithm in SVM technique to generate classification model and the testing dataset was used to evaluate the accuracy of the proposed model. Both datasets must be transformed to SVM format and gotten through scaling process in order to avoid attributes in greater numeric ranges dominating those in smaller ranges. Besides that, this process also wants to avoid numerical data representation difficulties during calculation. The recommended linearly scaling each attribute to the range [-1, +1] or [0, 1]. The sample of dataset used in the experiment phase is shown in Fig. 3.

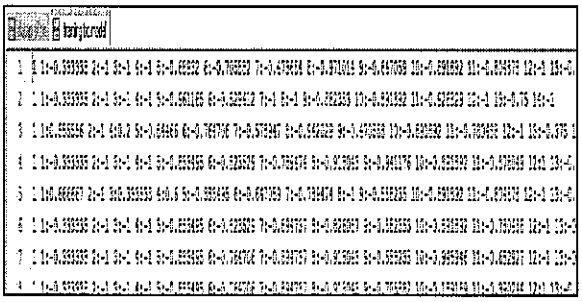


Figure 3. Sample of Dataset

In this study, SMO algorithm is used in the training process because it can handle optimization issues and can produce more accurate result efficiently. In this experiment, in order to determine the most accurate classification model, 10 randomize datasets (R1-R10) were used in the experiment by implementing 10 fold cross validation method for training and testing process. The accuracy of classification model for SVM techniques is represented by the average of accuracy for all randomized data. Table 2 shows the example of model analysis for 10 randomized datasets for 10 fold cross validation.

TABLE 2. SAMPLE OF MODEL ANALYSIS

Training: Testing	*R1 (%)*	*R2 (%)*	*R3 (%)*	*R4 (%)*	*R10 (%)*
10:90	66.67	66.67	74.10	70.38	51.85
20:80	60.38	64.15	54.72	56.60	52.83
30:70	55.13	57.70	57.69	51.28	56.41
40:60	64.15	55.66	47.17	50.00	55.66
50:50	58.65	57.14	60.15	46.62	55.64
60:40	52.20	50.31	54.09	40.88	57.23
70:30	43.32	48.13	56.68	51.34	47.60
80:20	46.70	50.00	47.64	44.81	50.00
90:10	38.24	35.29	41.60	51.68	36.13

ISBN: 978-1-941968-02-4 ©2014 SDIWC

The accuracy of the proposed model is based on the percentage of evaluation test data (untrained data) that are correctly classified.

4 RESULTS AND DISCUSSION

The accuracy of proposed SVM classification model for both algorithms (SOM and Grid Search) was determined based on the 10 fold cross validation for ten randomized data. Therefore, the accuracy of model for all randomized data is shown in Table 3.

TABLE 3. THE ACCURACY OF MODEL

No	Training	Testing	Accuracy
1	90	10	60.0
2	80	20	95.0
3	70	30	35.0
4	60	40	45.0
5	50	50	60.0
6	40	60	70.0
7	30	70	27.2
8	20	80	68.8
9	10	90	51.1

The result shows the highest accuracy is 95% for 80:20 training and testing model and the average of accuracy for 10 randomized datasets is 56.9% that slightly low (<80%). In the classification process, especially in model construction, the accuracy of model should be higher or acceptable enough in order to produce a good model from the selected algorithm before it can be applied to the actual data for classification or prediction.

In this study, there are some probable reasons regarding the accuracy of the model produced by SMO algorithm. The selection of attributes could be considered as one of the reason that affecting the accuracy of the model. The attribute selection process can be used to determine the importance of attribute by using some common methods such as Genetic algorithm, Boolean reasoning and others. Besides that, the number of datasets used in this experiment probably not enough to represent the whole talent knowledge and should consider in future work. Then, in data pre-processing especially in handling missing value or missing data the method used could be revising in order to produce good distribution in datasets. In model analysis phase, the higher accuracy of model (80:20) is selected that is produced 68 rules. Fig 4 shows the sample of classification rule generated using SVM classifier for the selected model that has been transformed to production rules.

```
if (years_of_service.getText(). equals("-
0.33") &&
(service_as_senior.getText().equals("-1")
&& (service_as_assiociate.getText().
equals("0.4")
&& (service_as_prof.getText().
equals("0.6") && (no_research.getText().
equals("-0.68 ")
&& (no_chief_research.getText().equals("-
0.70")
&& (national_journal.getText().
equals("0.47") &&
(s_award.getText().equals("-1")
&& (national_preceding.getText().
equals("-0.97")
&& (inter_journal.getText().equals("-
0.64") && (inter_preceding.getText(
).equals("-0.89")
&& (admin.getText().equals("-1")&&
(r_award.getText().equals("-0.5"))
{
        Prediction.setText("PROFESSOR");
}
```

Figure 4. Production Rules Representation

The generated rules can be embedded into a decision support system for employee's performance classification as a part of knowledge-based system component in the application. In model evaluation on untrained data using proposed prototype there are 30 untrained data used for this purpose. As a result, 19 data are correctly classified which is about 63.3% accuracy and it is considered acceptable for the datasets and it can be used in talent classification.

As an example of application, the employee achievement classification result is based on the

ISBN: 978-1-941968-02-4 ©2014 SDIWC

production rules transformed from the proposed SOM classification model that embedded in the system as shown in fig. 5.

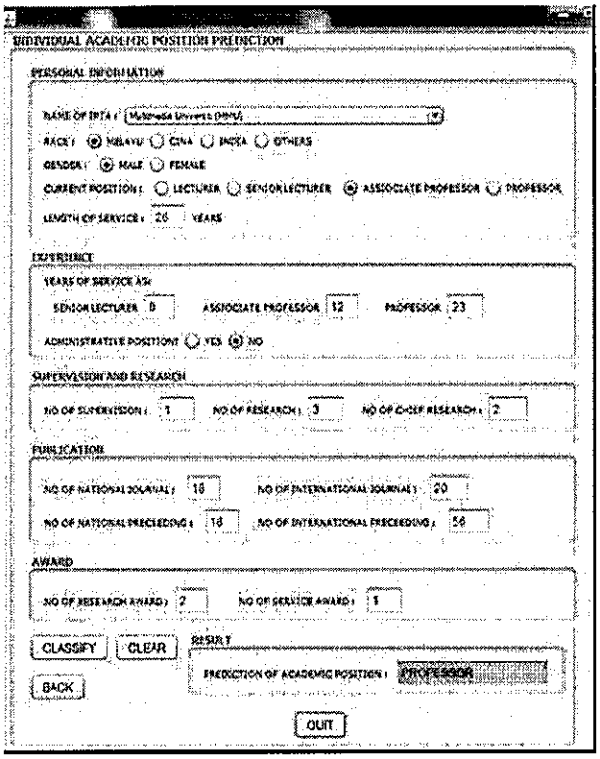


Figure 5. Employee Achievement Analyses

5 CONCLUSIONS AND FUTURE WORK

In this study, SOM algorithm from SVM technique is proposed as a classification method for employee's achievement classification. As a result, the accuracy of model proposed by SOM algorithm is considered acceptable and it need some enhancements in order to produce higher accuracy. In future work, the accuracy of classification model can be enhanced by a comparative study conducted using other SVM algorithms such as Grid Search and Gabriel graph algorithms that focus on reducing the size of a training set. It would give a direction on which algorithms can produce better result in SVM algorithms. Thus, this algorithm will be used as an alternative method in constructing classification rules for future achievement of employee identification. Besides that, the attribute selection process should be conducted using any attribute reduction techniques to compare the accuracy of proposed model after reduction whether it affect the accuracy of model. As conclusion, the ability to obtain new understanding of SVM classification technique in human resource decision system leads to the imperative contribution in HR field especially in talent management activities.

6 REFERENCES

[1] R. Mathur, "Talent Acquisition: A Challenge For Human Resource Professionals," *Lachoo Management Journal,* vol. 1, pp. 103 - 109, 2010.

[2] J. W. Boudreau and P. M. Ramstad, "Talentship and the Evolution of Human Resource Management: From "Professional Practices" To "Strategic Talent Decision Science"," 2004.

[3] P. Cappelli, "Talent Management for the Twenty-First Century," *Harvard Business Review,* pp. 1-9, 2008.

[4] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Human Talent Prediction in HRM using C4.5 Classification Algorithm," *International Journal on Computer Science and Engineering,* vol. 02, pp. 2526 - 2534, 2010.

[5] J. Han and M. Kamber, *Data Mining : Concepts and Techniques.* San Francisco: Morgan Kaufmann Publisher, 2006.

[6] C. F. Chien and L. F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," *Expert Systems and Applications,* vol. 34, pp. 380-290, 2008.

[7] J. Ranjan, "Data Mining Techniques for better decisions in Human Resource Management Systems," *International Journal of Business Information Systems,* vol. 3, pp. 464 - 481, 2008.

[8] L. Sadath, "Data Mining: A Tool for Knowledge Management in Human Resource," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 2, pp. 154-159, 2013.

[9] S. Strohmeier and F. Piazza, "Domain Driven Data Mining in Human Resource Management: A Review," *Expert System with Applications,* vol. 40, pp. 2410-2420, 2013.

[10] S.Yasodha and P. S.Prakash, "Data Mining Classification Technique for Talent Management using SVM," presented at the International Conference on Computing, Electronics and Electrical Technologies, 2012.

[11] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, and Q. Yang, "Top 10 Algorithms in Data Mining," *Knowledge Information System,* vol. 14, pp. 1-34, 2008.

[12] K.-S. Shin, T. S. Lee, and H.-J. Kim, "An Application of Support Vector Machines in Bankruptcy Prediction Model," *Expert System with Applications,* vol. 28, pp. 127-135, 2005.

[13] C. Shang and D. Barnes, "Support Vector Machine-Based Classification of Rock Texture Images Aided by Efficient Feature Selection," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, Brisbane, QLD, 2012, pp. 1 - 8.

[14] X. Su, Z.-R. Zhan, and C.-P. Li, "Forecasting Regional Talent Demand Based on Support Vector Machine," in *International Conference on Machine Learning and Cybernetics (ICMLC)*, Xian, 2012, pp. 355 - 359

[15] E. Aksakala, M. Dagdevirena, E. Eraslanb, and I. Yukselc, "Personel Selection Based on Talent Management," *Procedia - Social and Behavioral Sciences,* pp. 68 - 72, 27 February 2013.

[16] H. Jantan, A. R. Hamdan, and Z. A. Othman, "Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application," presented at the World Academy of Science, Engineering and Technology, Penang,Malaysia, 2009.

[17] P. Cappelli. (2008, March) Talent Management for the Twenty-First Century. *Harvard Business Review.* 1 - 9. Available: http://www.hbr.org

[18] A. H. Brayfield and W. H. Crockett. (1955) Employee Attitudes vs Employee Performance. *Psychological Bulletin.* 396-424.

[19] M. Corn and N. A. Esmen, "Workplace Exposure Zone for Classification of Employee Exposures to Physical and Chemical Agents," *American Industries Hygiene Association Journal,* vol. 40, pp. 47-57, 1979.

[20] M. Chandrasekaran, M. Muralidhar, C. M. Krishna, and U. S. Dixit, "Application of Soft Computing Techniques in Machining Performance Prediction and Optimization : A Literature Review," *International Journal of Advance Manufacturing Technology,* vol. 46, pp. 445-464, 2010.

[21] C. Yan-jiang and F. Xiao-ming, "The Study of DEA Method in Performance Evaluation of Project Human Resource Management," in *International Conference on Management Science and Engineering (ICMSE)*, Melbourne, 2010, pp. 1022 - 1030

[22] H. Zhang, "Fuzzy Evaluation on the Performance of Human Resources Management of Commercial Banks Based on Improved Algorithm," in *2nd International Conference on Power Electronics and Intelligent Transportation System (PEITS)*, Shenzhen 2009, pp. 214 - 218.

[23] M. Saron, "Model Base On Human Resource System Using Classification Technique," *IJCEM International Journal of Computational Engineering & Management,* vol. 15, pp. 1 - 8, 2012.

[24] S. Yasodha and P. S. Prakash, "Data Mining Classification Technique for Talent Management using SVM," in *International Conference on Computing, Electronics and Electrical Technologies (ICCEET)*, 2012, pp. 959-963.

[25] H. Jantan, A. R. Hamdan, Z. A. Othman, and M. Puteh, "Applying Data Mining Classification Techniques for Employee's Performance Prediction," in *Knowledge Management 5th International Conference (KMICe2010)*, Kuala Terengganu, Terengganu Malaysia, 2010, pp. 645-652.

[26] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy comsumption : A comparison of regression analysis, decision tree and nerural networks," *Energy,* vol. 32, pp. 1761 - 1768, 2007.

[27] Hua Hu, Jing Ye, and Chunlai Chai, "A Talent Classification Method Based on SVM," in *International Symposium on Intelligent Ubiquitous Computing and Education*, Chengdu, China, 2009, pp. 160-163.

[28] S. Pradhan, K. Hacioglu, V. Krugler, W. Ward, J. H. Martin, and D. Jurafsky, "Support Vector Learning for Semantic Argument Classification," *Machine Learning,* vol. 60, pp. 11 - 39, 02 Jun 2005.

[29] Y. Zhang, G. Shi, and K. Wang, "A SVM-HMM Based Online Classifier for Handwritten Chemical Symbols," presented at the International Conference on Pattern Recognition, 2010.

[30] Benoit Magnin, Lilia Mesrob, Serge Kinkingnehun, Melanie Pelegrini-Issac, Olivier Colliot, Marie Sarazin, Bruno Dubois, Stephane Lehericy, and H. Benali, "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI," *Neuroradiology,* vol. 51, pp. 73 - 83, February 2009.

[31] W. Zheng and Q. Ye, "Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm," in *Third International Symposium on Intelligent Information Technology Application, IITA 2009*, Nanchang, 2009, pp. 335 - 338

[32] S. Pradhah, K. Hacioglu, and V. Krugler, "Support Vector Learning for Sementic Argument

Classification," *Machine Learning,* vol. 60, pp. 12-38, 2005.

[33] Qiangwei Wang, Boyang Li, and Jinglu Hu, "Feature Selection for Human Resource Selection Based on Affinity Propogation and SVM Sensitivity Analysis," in *World Congress on Nature & Biologically Inspired Computing (NaBIC 2009)*, Coimbatore, India, 2009, pp. 31-36.

[34] A. Barbero and J. e. R. Dorronsoro, "Momentum Sequential Minimal Optimization: An accelerated method for Support Vector Machine training " in *The 2011 International Joint Conference on Neural Networks (IJCNN)*, San Jose, 2011, pp. 370 - 377.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Construction of Subject-independent Brain Decoders for Human FMRI with Deep Learning

Sotetsu Koyamada[1,2], Yumi Shikauchi[1,2], Ken Nakae[1] and Shin Ishii[1,2]
1. Graduate School of Informatics, Kyoto University, Kyoto, Japan
2. ATR Cognitive Mechanisms Laboratories, Kyoto, Japan
koyamada-s@sys.i.kyoto-u.ac.jp, ishii@i.kyoto-u.ac.jp

ABSTRACT

Brain decoding, to decode a stimulus given to or a mental state of human participants from measurable brain activities by means of machine learning techniques, has made a great success in recent years. Due to large variation of brain activities between individuals, however, previous brain decoding studies mostly put focus on developing an individual-specific decoder. For making brain decoding more applicable for practical use, in this study, we explored to build an individual-independent decoder with a large-scale functional magnetic resonance imaging (fMRI) database. We constructed the decoder by deep neural network learning, which is the most successful technique recently developed in the field of data mining. Our decoder achieved the higher decoding accuracy than other baseline methods like support vector machine (SVM). Furthermore, increasing the number of subjects for training led to higher decoding accuracy, as expected. These results show that the deep neural networks trained by large-scale fMRI databases are useful for construction of individual-independent decoders and for their applications for practical use.

KEYWORDS

FMRI, brain decoding, brain machine interface (BMI), subject-independent decoding, deep learning

1 INTRODUCTION

Brain decoding is a technology to read out (decode) a stimulus given to or a mental state of human participants from measurable brain activities, which has potential applications in neuroscience-based engineering, such as brain machine interface, neuro rehabilitation, and even therapy of mental disorders. Brain decoding is usually based on machine learning, especially, supervised learning framework; the decoder is trained to associate brain activities as its input and stimuli or mental states as its output. Because brain activities are very different between individuals, previous brain decoding studies mostly focused on construction of subject-dependent decoder for each subject (e.g., [1], [2], [3], [4]). In practical situations of applying brain decoding, however, it may be difficult to collect sufficient data for training subject-dependent decoders for various reasons; especially in the scenario of BMI, the subjects could be disabled, then they may not be able to perform a number of task sessions to collect sufficient amount of data. When considering practical brain decoding technology, construction of subject-independent decoders based on extraction of subject-independent features inside has been highly demanded. Here, subject-independent decoders are required to read out the brain activities of an unseen subject whose data have never been used for training the decoders.

With an interest in building subject-independent decodes, in this study, we applied deep neural network learning to a large fMRI dataset which includes many subjects' data when performing various kinds of cognitive tasks. In particular, we used the Human Connectome Project (HCP) dataset [5], which is one of the largest public-available fMRI databases. HCP includes fMRI data of over 500 subjects when they are performing seven kinds

ISBN: 978-1-941968-02-4 ©2014 SDIWC

of cognitive tasks. The deep neural network leaning has the potential to make the best use of this 'big data'; it recently attracts much attention because of its high classification performance in various artificial intelligence issues, like image recognition, speech recognition, and so on [6], [7]. Very recently, some studies applied the deep learning technique to analyses of fMRI data; Plis et al., [8] compared schizophrenia patients and healthy controls, and Hatakeyama et al., [9] presented an application to voxel-wise decoding of hand motions. However, there has been no study that used the deep learning technique for subject-independent decoding of cognitive tasks, especially with the help of big data. To our best knowledge, this is the first study, so would be important for allowing the brain decoding technology to be applicable to many practical situations.

2 METHODS

2.1 Data Acquisition and Preprocessing

In this study, we used the preprocessed task-evoked fMRI data registered in the HCP Q3 fMRI database [5], [10]. The HCP dataset is one of the largest open databases, covering fMRI data during various types of cognitive tasks. Here, we briefly explain key data specifications and preprocessing procedure. For more details, see HCP Q3 Release Reference Manual(www.humanconnectome.org/documentation/Q3).

FMRI data were acquired from eighty healthy and unrelated adult subjects, by a Siemens 3T Skyra, with TR = 720 ms, TE = 33.1 ms, flip angle 52°, FOV = 208×180 mm, 72 slices, 2.0×2.0 mm in plane resolution. Our fMRI data have been applied by low level pre-processing: removal of spatial artifacts and distortions, within-subject cross-modal registrations, reduction of the bias field, and normalization to standard space [10]. To the preprocessed fMRI data, we applied voxel-wise z-score transformation, followed by averaging over each anatomical region of interests (aROI) to obtain robust features against the large inter-subject variability of brain activities. AROIs were determined by the automated anatomical labeling method [11] for each subject, which utilized the anatomical predefinition in terms of templates in the WFU PickAtlas [12]. After these preprocesses, the dimension per fMRI scan was 116.

Each of the eighty subjects performed all of seven tasks: emotion, gambling, language, motor, relational, social and working memory (WM), for two runs, and each cognitive task continued for different time duration (see Table 1). The experimental design of each task is summarized below. See Barch et al., [13] for more details.

1. **Emotion**: this task was a modified version of Hariri et al., [14]. Participants were required to match one of two simultaneously presented images with a target image (angry face or fearful face).
2. **Gambling**: participants guessed the number on a card in order to win or lose money. See Delgado et al., [15] for more details.
3. **Language**: after listening to a brief story, participants were asked a two-alternative forced choice question about the topic of the story. See Binder et al., [16] for more details.

Table 1. Number of scans per run and run duration (min).

	Emotion	Gambling	Language	Motor	Relational	Social	WM
Scans	176	253	316	284	232	274	405
Duration	2:16	3:12	3:57	3:34	2:56	3:27	5:01

ISBN: 978-1-941968-02-4 ©2014 SDIWC

4. **Motor**: participants were requested to move one of five body parts (left or right finger, left or right toe, or tongue) instructed by a visual cue [17].
5. **Relational**: this task was a modified version of Smith et al., [18]. Participants answered a second-order question between two pairs of objects, whether or not these pairs share the mismatch dimension (texture or shape) across the pair.
6. **Social**: participants were asked if objects in video clips interacted in some way or not. These videos were taken from either Castelli et al., [19] or Wheatley et al., [20].
7. **Working memory**: two-back working memory task and zero-back working memory task with four different types of picture stimuli (places, tools, faces or body parts).

2.2 Decoding with Deep Learning

The objective of deep neural network leaning was to acquire the input-output relationship with the input being the fMRI signals and the output being their labeled task classes, i.e., the category of cognitive task performed by the participants. For example, each fMRI scan during the participant performed the emotion task was labeled as 'emotion class'. Then, the deep neural network was required to solve the classification problem into seven classes according to the supervised learning framework. As shown in Table 1, the number of scans in a single run was different between the tasks. To avoid harmful influence stemming from this difference in the data number, we resized the number of samples by randomly sub-sampling for each participant, hence the number of samples per run became 176, common for all tasks. This number 176 was the same as the smallest scan number per run among the seven tasks. Hence, the total sample number in the dataset was 176×2×80 for each class. The architecture and learning method of deep neural networks used in this study are similar to those previously used in the MNIST classification experiments of Hinton et al., [21].

A neural network was configured as a feed-forward network incorporating L hidden layers. The internal potential of the i-th unit in the l-th hidden layer $a_i^{(l)}(l = 1, \cdots L)$ is given as a weighted summation of its inputs:

$$a_i^{(l)} = \sum_{j=1}^{n_{l-1}} w_{ij}^{(l)} z_j^{(l-1)} + b_i^{(l)} \quad (1)$$

where $w_{ij}^{(l)}$ and $b_i^{(l)}$ are a weight and a bias. n_l is the number of units in the l-th hidden layer, which was set at $n_l = 500$ for any $l > 0$. $\boldsymbol{z}^{(0)}$ denotes the input vector $\boldsymbol{x}$ to the network, hence n_0 equals to the input's dimension $d(=116)$. $\boldsymbol{z}^{(l)} = \left(z_1^{(l)}, \cdots, z_{n_l}^{(l)}\right)^\top$ represents the output of the l-th hidden layer and is given by applying a nonlinear activation function f to the internal potential as

$$z_i^{(l)} = f\left(a_i^{(l)}\right). \quad (2)$$

Here, ReLU [22], a piecewise linear function $\max(0, x)$, was used for the activation function f. Usage of ReLU for the activation function has a couple of advantages; its piecewise linearity can save the computational cost to calculate its derivative, and its non-saturating character prevents the learning algorithm from halting due to gradient vanishing of nonlinear activation functions.

The last hidden layer was connected to the softmax (output) layer, so that the output from the k-th unit of the output layer was interpreted as the posterior probability of class k, given by

$$P(Y = k|\boldsymbol{x}, \boldsymbol{W}) = \frac{\exp\left(\sum_{j=1}^{n_L} w_{kj} z_j^{(L)} + b_k\right)}{\sum_{k'=1}^{K} \exp\left(\sum_{j=1}^{n_L} w_{k'j} z_j^{(L)} + b_{k'}\right)} \quad (3)$$

ISBN: 978-1-941968-02-4 ©2014 SDIWC

where $K(=7)$ is the number of classes, and $\boldsymbol{W}$ denotes all the parameters (weights and biases). Y is a random variable signifying the class to which $\boldsymbol{x}$ belongs.

We used a negative log-likelihood as the cost function of the learning

$$L(\boldsymbol{W}) = -\sum_{n=1}^{N} \log P(Y_n = t_n | \boldsymbol{x}_n, \boldsymbol{W}) \qquad (4)$$

where $\{(\boldsymbol{x}_1, t_1), \cdots, (\boldsymbol{x}_N, t_N)\}$ constituted the given dataset. $t \in \{1, \cdots, K\}$ denotes a class label. To minimize the above cost function, minibatch stochastic gradient descent (MSGD) with a momentum was introduced so that the stochastic gradient descent was performed every 100 samples:

$$\boldsymbol{W}_t = \boldsymbol{W}_{t-1} + \boldsymbol{v}_t \qquad (5)$$

$$\boldsymbol{v}_t = p_t \boldsymbol{v}_{t-1} - (1 - p_t)\eta_t \left.\frac{\partial L'(\boldsymbol{W})}{\partial \boldsymbol{W}}\right|_{\boldsymbol{W}=\boldsymbol{W}_{t-1}} \qquad (6)$$

where L' is the cost function for the cached subset of 100 samples in the minibatch, and η_t and p_t are the learning rate and the momentum rate, respectively. The learning rate η_t started with η_0, then was exponentially decreased as $\eta_t = r\eta_{t-1}$. The momentum rate p_t was increased linearly from p_0 to $p_{100} = 0.99$; after 100 times updates, p_t was fixed at p_{100}. When searching for appropriate values of the hyper-parameters (η_0, r, p_0, l), we used random search rather than grid search [23], in which η_0, r, p_0 and l were randomly sampled from their individual uniform distributions on the intervals $[1.0, 20.0]$, $[0.95, 0.9999]$, $[0.4, 0.6]$ and $[3.0, 20.0]$ respectively. The best parameters were chosen among 9 combinations of (η_0, r, p_0, l).

Each weight was initialized as a small value randomly sampled from a zero-mean normal distribution with the standard deviation of 0.01, and biases were initialized to zero. During learning, the weight vector of each hidden unit $\left(w_{i1}^{(l)}, \cdots, w_{in_{l-1}}^{(l)}\right)^{\mathsf{T}}$ was not allowed to make its L_2 norm larger than a fixed positive constant m. If the L_2 norm of the weight vector got larger than m after each update, it was simply divided by the norm and then multiplied by m. This upper bound setting of the norm enabled the initial learning rate to be fairly large, by which we expected accelerated learning. Early stopping was also adopted. If the decoding accuracy for the validation dataset did not increase for 200 learning epochs, then learning was terminated. Even if the early stopping did not occur, the whole learning procedure was terminated after 5000 learning epochs.

For avoiding over-fitting, we used the dropout technique [21]. During training, the activity $z_i^{(l)}$ was randomly replaced by 0 with probability p. We set $p = 0.5$ for hidden units and 0.2 for inputs. This dropping out of activities plays a role of regularization and is expected to prevent the decoder from acquiring subject-specific features. When testing the trained neural network, on the other hand, all the nodes were activated, but their weights were multiplied by $1 - p$, to make the mean activity level of each network element consistent between the training phase and the test phase (see Fig. 1).

The trained neural network was tested by unseen data. Since in this study we expect the deep neural network can extract subject-independent features based on training from the large-scale fMRI database, we examined subject-transfer decoding performance. In specific, we executed 8-fold cross validation, or equivalently, *leave-10-subjects-out* cross validation: the whole dataset of 80 subjects were repeatedly separated into a training dataset of 70 subjects and a test dataset of the remaining 10 subjects. In addition, 10 subjects were randomly taken from the training dataset of 70 subjects to construct a validation dataset,

ISBN: 978-1-941968-02-4 ©2014 SDIWC

which was in turn used for determining hyper-parameters and early stopping criterion.

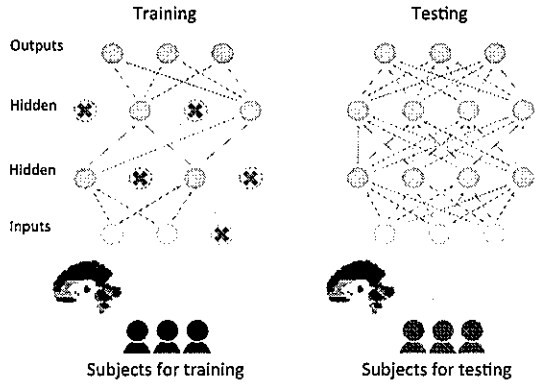


Figure 1. Training and testing a deep neural network in subject transfer decoding. (Left) When training a deep neural network, we used a dropout technique for regularizing the neural network learning; when learning a single example in the training dataset, a half of hidden units and 20% of input units were ignored without emitting their outputs to the network or learning. (Right) When testing the trained deep neural network, all the units were activated, whereas the weights of the hidden units were lowered into their halves to be balanced with the training situations. Due to the setting of subject transfer decoding, the test data were from the test subjects other than the training subjects included in the training dataset.

3 RESULTS

First, we compared the decoding accuracy of the deep neural networks with those of other baseline methods. We trained three neural networks with one, two and three hidden layers, and another network with two hidden layers ($L = 2$) without dropout and with sigmoid activation functions; the last one was to know the improvement achieved by the dropout and ReLU. As baseline methods, logistic (softmax) regression, which corresponds to 0-hidden layer neural network, and SVMs with linear kernel and RBF kernel were trained; SVMs were configured to be one-versus-the-rest multi-class classifiers. Deep neural networks and logistic regression were implemented by Theano [24], and we used a learning kit 'scikit-learn' [25] for SVM. Hyper-parameter was determined such to maximize the decoding accuracy for validation dataset. Since the number of hyper-parameters was different between the competitor methods, nine sets of hyper-parameter values were examined and the best set was selected in each method. The logistic regression had a learning rate hyper-parameter, and the linear-kernel SVM had a regularization hyper-parameter C. The RBF-kernel SVM had a couple of hyper-parameters, a regularization parameter C and a kernel coefficient γ; in this case, we used grid search for the best combination of C and γ over $3\times3 = 9$ patterns. The results are summarized in Fig. 2. The three neural networks with dropout and ReLU activation showed reasonably good decoding accuracy of more than 45 %. Their decoding accuracies were higher than those by the other baseline methods and the chance level of 14.29%(=100%/7). Especially, the deep neural network with two hidden layers exhibited the best decoding accuracy of 48. 24 %. Linear methods, the logistic regression and the linear-kernel SVM, showed poor decoding accuracies comparable to the chance level, clearly showing the advantage of the non-linear decoding methods. These results suggest that the deep neural networks were more effective in extracting subject-independent features within its non-linear architecture, leading to higher subject-transfer decoding accuracies.

Second, we examined how the subject-transfer decoding performance behaved when the number of subjects included in the training dataset was increased from 20 to 70. In this evaluation, we compared the deep neural network with two hidden layers ($L = 2$) and the RBF-kernel SVM. For each number of the training subjects, we took 10 subjects to construct a validation dataset to tune the hyper-parameters. The results are displayed in Fig. 3.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

As the number of training subjects increased, the performance of the deep neural network also increased as expected; this would be owing to the improvement of subject-independent features extracted by the network as the number of training subjects increased. Although the same character was observed in the SVM learning, its performance was consistently inferior to that by the deep neural network. This result implies that the subject-transfer decoding would become more practical if we can access to larger brain signal databases including even larger number of subjects. Such good usability of 'big data' was naturally incorporated by the non-linear learning scheme based on deep neural networks.

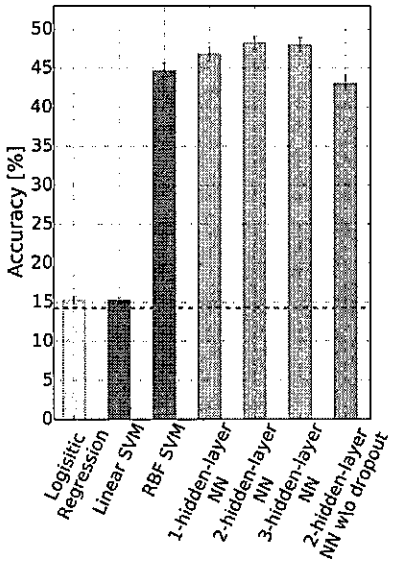


Figure 2. Comparison of decoding accuracy between the deep neural networks and other decoding methods; they are logistic regression (yellow), which corresponds to 0-hidden layer neural network, SVMs with linear kernel and RBF kernel (blue), and deep neural networks (NN) with one, two, and three hidden layers (light red). The training of the neural networks were performed with ReLU and dropout. For reference, decoding accuracy of NN without dropout and with sigmoid activation functions is also shown (magenta). Each error bar is the 95% confidence interval of the decoding accuracy. A red dotted line denotes the chance level.

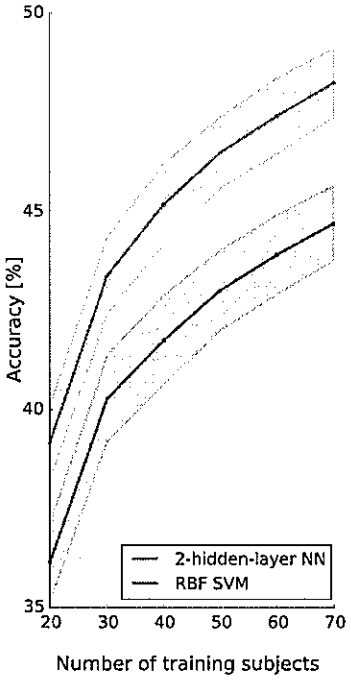


Figure 3. The accuracy (in terms of the subject transfer decoding accuracy) against the number of subjects included in the training dataset. The red and blue curves depict the accuracy of deep neural network (number of hidden layers, $L = 2$) and SVM with RBF kernel, respectively. The red and blue shaded regions correspond to the 95% confidence intervals.

Fig. 4 shows the time-series of the decoder's prediction by the deep neural network with two hidden layers, which showed the best subject-transfer decoding accuracy in the 8-fold cross-validation. We show the average of the decoder's outputs, corresponding to the average posterior probability that the task is belonging to each of the seven classes, along the time profile of sessions of each task class. This result shows that some cognitive tasks (e.g., language) were relatively easy to discriminate, but some others (e.g., WM) were somehow difficult. Such discriminability would be dependent on the distance in the feature space between task classes. Moreover, we observe some zig-zag patterns in the decoder's class

ISBN: 978-1-941968-02-4 ©2014 SDIWC

prediction, as typically seen in panels (d), (e) and (f). These patterns occurred because there were resting states between two subsequent task sessions. That is, the decoder, which is nothing detected difference in the brain activities between task periods and resting periods in a data-driven manner.

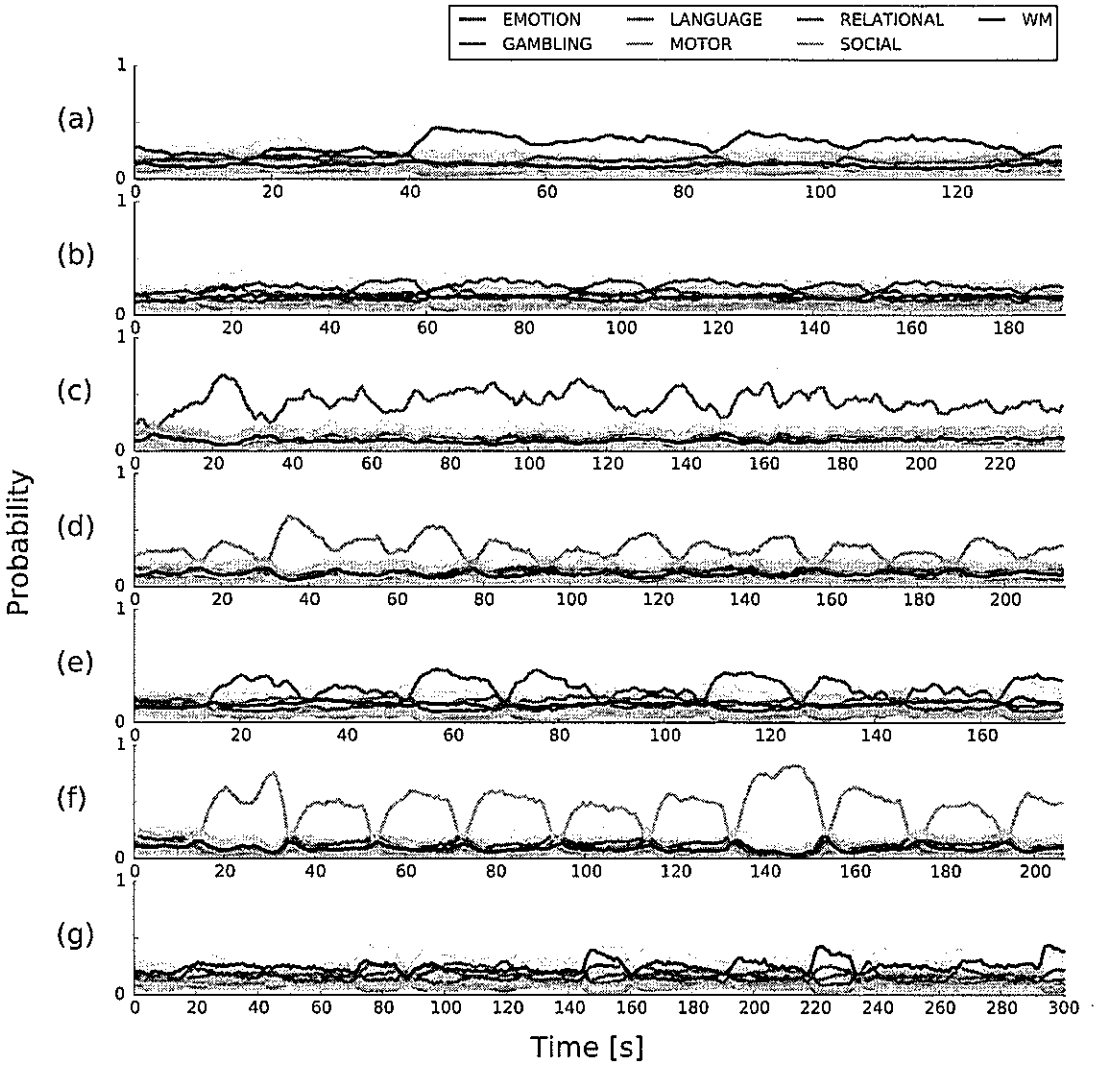


Figure 4. The time-series of the class prediction by the deep neural network with two hidden layers. We show the average of the decoder's outputs, corresponding to the average posterior probability that the task is belonging to each of the seven classes, along the time profile of sessions of each task class: (a) emotion, (b) gambling, (c) language, (d) motor, (e) relational, (f) social, or (g) WM. Each single time-series in each panel corresponds to the decoder's output representing the respective posterior probability for each of the seven classes, whose color is defined in the inset. The shaded color denotes the standard deviation.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

4 CONCLUSION

In this study, we proposed to use deep neural network learning for constructing task classification decoders trained by a large dataset from a public fMRI database. The trained decoders were also available for subject-transfer decoding. As a result, our approach based on deep learning achieved the higher decoding accuracy than other baseline methods, and got even improved as the number of training subjects increased. We thus concluded the deep neural network learning was ready for obtaining subject-independent non-linear features from a 'big-data' of brain activities, and then for applying to subject-transfer decoding, which is an important methodology for making the brain-machine-interface more practical in realistic situations.

5 ACKNOWLEDGEMENT

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

This research was supported in part by the Ministry of Internal Affairs and Communications, Japan, under a contract "Novel and innovative R&D making use of brain structures" and by JSPS KAKENHI (No. 24300114).

REFERENCES

[1] J. V Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini, "Distributed and overlapping representations of faces and objects in ventral temporal cortex," *Science*, vol. 293, no. 5539, pp. 2425–2430, 2001.

[2] D. D. Cox and R. L. Savoy, "Functional magnetic resonance imaging (fMRI) 'brain reading': detecting and classifying distributed patterns of fMRI activity in human visual cortex," *Neuroimage*, vol. 19, no. 2, pp. 261–270, 2003.

[3] S. Nishimoto, A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant, "Reconstructing visual experiences from brain activity evoked by natural movies," *Current Biology*, vol. 21, no. 19, pp. 1641–1646, 2011.

[4] T. Horikawa, M. Tamaki, Y. Miyawaki, and Y. Kamitani, "Neural decoding of visual imagery during sleep," *Science*, vol. 340, no. 6132, pp. 639–642, 2013.

[5] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. J. Behrens, E. Yacoub, and K. Ugurbil, "The WU-Minn Human Connectome Project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[7] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 437–440, 2011.

[8] S. M. Plis, D. R. Hjelm, R. Salakhutdinov, E. a Allen, H. J. Bockholt, J. D. Long, H. J. Johnson, J. S. Paulsen, J. a Turner, and V. D. Calhoun, "Deep learning for neuroimaging: a validation study," *Frontiers in Neuroscience*, vol. 8, p. 229, Jan. 2014.

[9] Y. Hatakeyama, S. Yoshida, H. Kataoka, and Y. Okuhara, "Multi-voxel pattern analysis of fmri based on deep learning methods," in *Soft Computing in Big Data Processing*, vol. 271, pp. 29–38, 2014.

[10] M. F. Glasser, S. N. Sotiropoulos, J. A. Wilson, T. S. Coalson, B. Fischl, J. L. Andersson, J. Xu, S. Jbabdi, M. Webster, J. R. Polimeni, D. C. Van Essen, and M. Jenkinson, "The minimal preprocessing pipelines for the Human Connectome Project," *Neuroimage*, vol. 80, pp. 105–124, 2013.

[11] N. Tzourio-Mazoyer and B. Landeau, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *Neuroimage*, vol. 15, no. 1, pp. 273–289, 2002.

[12] J. a. Maldjian, P. J. Laurienti, R. a. Kraft, and J. H. Burdette, "An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets," *Neuroimage*, vol. 19, no. 3, pp. 1233–1239, 2003.

[13] D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, D. Nolan, E. Bryant, T. Hartley, O. Footer, J. M. Bjork, R. Poldrack, S. Smith, H. Johansen-Berg, A. Z. Snyder, and D. C. Van Essen, "Function in the human connectome: task-fMRI and individual differences in behavior," *Neuroimage*, vol. 80, pp. 169–189, 2013.

[14] A. R. Hariri, A. Tessitore, V. S. Mattay, F. Fera, and D. R. Weinberger, "The Amygdala Response to Emotional Stimuli: A Comparison of Faces and Scenes," *Neuroimage*, vol. 17, no. 1, pp. 317–323, 2002.

[15] M. R. Delgado, L. E. Nystrom, C. Fissell, D. C. Noll, and J. A. Fiez, "Tracking the hemodynamic responses to reward and punishment in the striatum," *Journal of Neurophysiology*, vol. 84, no. 6, pp. 3072–3077, 2000.

[16] J. R. Binder, W. L. Gross, J. B. Allendorfer, L. Bonilha, J. Chapin, J. C. Edwards, T. J. Grabowski, J. T. Langfitt, D. W. Loring, M. J. Lowe, K. Koenig, P. S. Morgan, J. G. Ojemann, C. Rorden, J. P. Szaflarski, M. E. Tivarus, and K. E. Weaver, "Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study," *Neuroimage*, vol. 54, no. 2, pp. 1465–1475, 2011.

[17] R. L. Buckner, F. M. Krienen, A. Castellanos, J. C. Diaz, and B. T. T. Yeo, "The organization of the human cerebellum estimated by intrinsic functional connectivity," *Journal of Neurophysiology*, vol. 106, no. 5, pp. 2322–2345, 2011.

[18] R. Smith, K. Keramatian, and K. Christoff, "Localizing the rostrolateral prefrontal cortex at the individual level," *Neuroimage*, vol. 36, no. 4, pp. 1387–1396, 2007.

[19] F. Castelli, F. Happé, U. Frith, and C. Frith, "Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns," *Neuroimage*, vol. 12, no. 3, pp. 314–325, 2000.

[20] T. Wheatley, S. C. Milleville, and A. Martin, "Understanding animate agents: distinct roles for the social network and mirror system," *Psychological. Science*, vol. 18, no. 6, pp. 469–474, 2007.

[21] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv preprint arXiv:1207.0580, 2012.

[22] K. Jarrett, K. Kavukcuoglu, M. A. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2146–2153, 2009.

[23] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *The Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[24] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-farley, and Y. Bengio, "Theano : A CPU and GPU Math Compiler in Python," in *Proceedings of the Python for scientific computing conference (SciPy)*, pp. 1–7, 2010.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2012.

Extraction of Automatic Search Result Records Using Content Density Algorithm Based on Node Similarity

Yasar Gozudeli*, Oktay Yildiz*, Hacer Karacan*, Mohammed R. Baker*, Ali Minnet**, Murat Kalender**, Ozcan Ozay**, M. Ali Akcayol*
*Gazi University, Faculty of Engineering, Department of Computer Engineering, Maltepe, Ankara, Turkey
**Huawei Technologies Co., Ltd., Istanbul, Turkey
ygozudeli@gmail.com, oyildiz@gazi.edu.tr, hkaracan@gazi.edu.tr, mr_baker@live.com, ali.minnet@huawei.com, murat.kalender@huawei.com, ozcan.ozay@huawei.com, akcayol@gazi.edu.tr

ABSTRACT

In this paper, a new method proposed for finding and extracting the SRRs. The method first detects content dense nodes on HTML DOM and then extracts SRRs to suggest a list of candidate HTML DOM nodes for a given single research result Web page instance. Afterwards an evaluation algorithm has been applied to the candidate list to find the best solution without any human interaction and manual process. Experimental results show that the proposed methods are successful for finding and extracting the SRRs.

KEYWORDS

Automatic Web extraction, deep Web, meta-search engines, search result extraction, tree similarity

1 INTRODUCTION

World Wide Web (WWW) is generally categorized into surface web and deep web with respect to ease of accessibility by agents. Surface web is conventional and has been crawled by all search engines since the invention of Internet. Academic researches on deep web have been expanded for last decade after the term "Deep Web" introduced at 2000 [11]. The number of web databases reached by agents is approximately 25 million pages at 2007 [5], [6].

The number of public accessible web data sources, Deep Web or Web databases, has increased continuously [4]. When compared to traditional databases, Web data is public and entirely accessible. However, web results obtained from search engines are quite difficult to extract. Generally, Web data results are represented in a markup language like Hyper Text Markup Language (HTML). Especially mash-ups and meta-search engines have to refine the Web results from the inorganic results like ads and some of other visual HTML elements.

Accessing and reusing a web search results in programmatic environment provides easiness in Web data mash-ups, meta-search engines, Web mining and Social web mining. However, there is no well-defined way to fetch the results of public search engines. Some of them may provide an Application Programming Interface (API) for applications to access, however this kind of support is not life time guaranteed and may be terminated abruptly. On the other hand, it is not a common practice for search engines and in fact the most of the search engines do not provide an API.

In meta-search engines and data mash-up systems, the result list gathered from public web search engines are called Search Result Records (SRR). Many studies have been proposed to extract web results automatically [12], [13], [14]. Two Popular categories of approaches in automatic deep Web extraction

ISBN: 978-1-941968-02-4 ©2014 SDIWC

are wrapper induction and automatic template generation. In addition, hybrid approach can be mentioned as another category.

A wrapper is used for web extracting systems in the wrapper induction applications. Wrapper is referred to as different functional ability in different studies [9]. However, it is defined as general meaning of any predefined code or symbolic definition used for web result extraction in this paper. Wrapper induction based approaches generally follow three steps: (1) wrapper generation, (2) wrapper execution and (3) wrapper maintenance. A wrapper must be defined or generated prior to usage by a developer, a user or software. After that, Web SRR can be extracted by executing the wrapper. Since Web resource templates change from time to time, wrapper has to be adjusted accordingly.

Another popular Web extraction category is automatic template generation [3]. This approach assumes that HTML result pages are generated by server side programs and results are located in common patterns. Automatic template generation approaches generally use tree structures to find repeating patterns [10]. While some researchers focus on visual items [4], [7] and use visual block trees; some other researchers uses HTML tag trees [8], [9].

A Web page containing web search results usually has a regular pattern to show the results but sometimes significantly different results may be found within the same result list, as shown in Figure 1. We call "irregular SRR" to this kind of SRRs. Irregular SRRs complicates detection of the repeating patterns. In addition, some of SRRs may contain HTML rich content and this introduces additional challenge to find the exact position of parent HTML nodes of SRRs.

In this study, a new approach has been developed for SRR extraction without any user interaction and manual process.

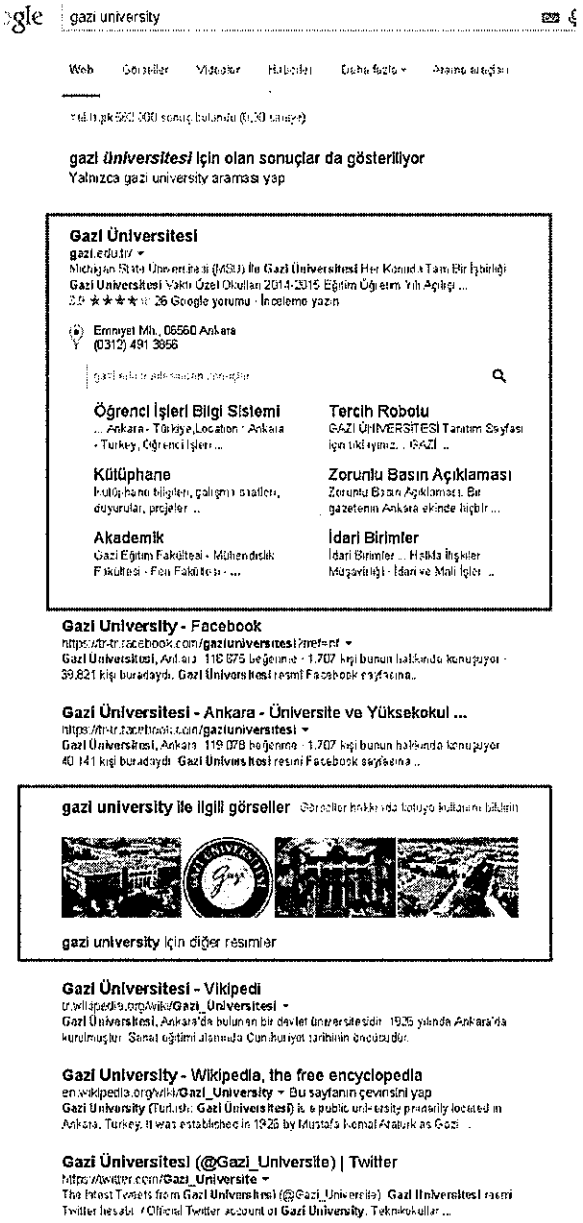


Figure 1. A sample Google search result containing some irregular SRRs.

Compared to existing approaches, this study has three contributions.

- A new method has been proposed for automatically detecting SRRs with a single search result page instance.
- Irregular SRRs are included in extraction process.

- The method is evaluated on a common Web databases and large variety of search results.

This paper is structured as follows. Related work is discussed in Section 2. In addition to describing approach, Section 3 also discusses the developed content density algorithm. Section 4 presents the test environment and the last section concludes the study.

2 RELATED WORKS

There have been many studies on Web result extraction in the literature. [12], [13] and [14] presents different methods and classifications of the main Web result extraction approaches.

The earliest studies have manual approaches based on wrapper induction in which a manually labeled web page instance is used for learning Web result extraction.

Later the development focused on various semi-automatic and automatic approaches. Semi-automatic methods can be classified into string based techniques and tree-based techniques. Wien [17] and Stalker [18] are examples of string based techniques. Web results are assumed as a flat-sequence of strings and delimiter sections are determined by the help of manually labeled training web result documents in these studies. W4F [15] and Wrapper [16] parse web documents into hierarchical trees (Document Object Model (DOM) trees) rather than flat-string. Afterwards, with the help of labeled training instances, a set of delimiter based rules is generated. Since string based and tree-based semi-automatic methods need human intervention, they are not appropriate for today's huge amount of Web data extraction processes.

In order to deal with the scale of deep Web, recent studies have advanced the automatic approaches [2]. IEPAD [19], MDR [20], RoadRunner [3], EXALG [21], DEPTA [22], Tag of Path [8] are some examples of studies. Automatic extraction approaches can be categorized according to their input data requirement before the operation.

IEPAD and MDR are focused on extracting Web results from only one Web page. While IEPAD identifies repeating substrings as tokens, MDR uses aggressive approach based on similarity match between two segments.

RoadRunner, EXALG and DEPTA use more than one page to extract the web results. RoadRunner, takes an initial web page as template and adjusts it for other pages, accordingly. EXALG assumes that any repeating tokens similar in different web pages are web results. DEPTA uses HTML DOM based partial tree alignment for finding the Web result section.

In the complex networks literature, "Estrada index" is used as a measure of centrality and computes importance values for each node [24], [25], [26]. Although node importance values can be used to find similar nodes; Estrada Index is not aware of tree node structure, which is essential in HTML tree processing context.

SRRs detection process has been categorized as visual tree processing approaches and HTML node processing. In this paper HTML node processing methods have been used. Tag of Paths [8] and Ranking XPaths [9] are most recent studies in this category. While Tag of Paths uses a vector representation to find the repeating HTML tags, Ranking XPaths approach uses the rank of XPath queries for the same purpose.

Ranking XPath assumes all of the SRRs are similar and assumes that there should be no outliers.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

3 METHODOLOGY

Main goal of this study is to create a new algorithm to suggest a list of candidate HTML DOM nodes for a given single research result Web page instance. Afterwards an evaluation algorithm is applied on the candidate list to find the best solution without any human interaction and manual process.

Thanks to the advances in information retrieving techniques, different irregular SRRs have been represented in search results as shown in Figure 2.

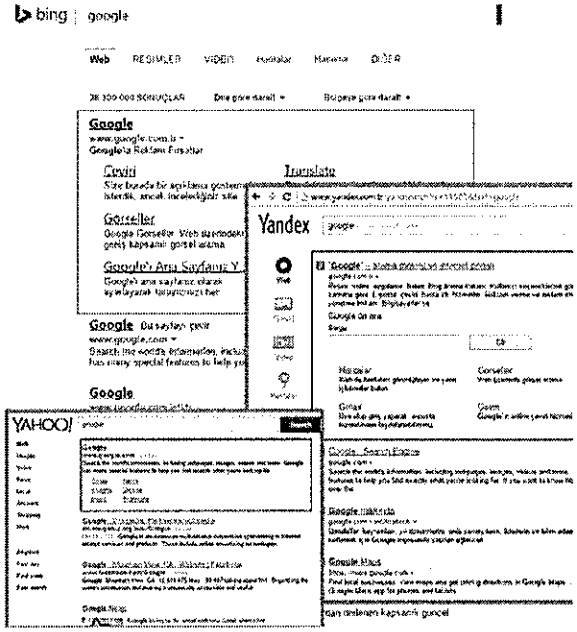


Figure 2. Irregular SRRs are very common

The most of well-designed search result page dedicate the major part of content to search results. For this reason, SRRs have to be located around the dense content regions. In [4], the threshold value of a data region ratio to the whole page content region is taken as 0.4. Although, it works well for much more of search result pages, the experimental results show that when search engines show more advertisement within the SRR page, a better result can be obtained by using 0.2 as threshold ratio.

The proposed approach has five steps:

- Generate candidate nodes by calculating content size factor (CSF) and content density for each node.
- Distinguish regular nodes and irregular nodes using the value of content density for all candidate nodes.
- Reduce tree complexity in SRRs by removing certain visual HTML tags like <b> and .
- Only regular nodes are evaluated by tree edit distance algorithm.
- Best scored nodes which located under the same node are returned as SRR list.

Content size factor shows the size of content and can be calculated for text and image differently. While, length and font size are used for the text content, width and height are used for image content to calculate CSF. CSF based content density has been calculated for each candidate node by the equation as shown below:

$$Content\ density = \frac{CSF}{Sub\ node\ count} \quad (1)$$

In Table-1, first and second lines have potential to be irregular SRRs. The last three rows seem to be regular SRRs.

Table 1. A sample of candidate node for a sample search query

Node Path	# of sub nodes	Content size factor	Content density
table/tbody/tr/td/ div/ol/li	3	4253	1417
table/tbody/tr/td/ div/ol/li	1	2548	2548
table/tbody/tr/td/ div/ol/li	2	1186	593
table/tbody/tr/td/ div/ol/li	2	1340	670
table/tbody/tr/td/ div/ol/li	2	1441	720

ISBN: 978-1-941968-02-4 ©2014 SDIWC

In order to evaluate the candidate nodes, similarity score is calculated for only the regular SRRs, rather than all SRRs.

Braces organized representation and tree edit distance algorithm [23] is used for visual representation during the evaluation process as shown Figure 3.

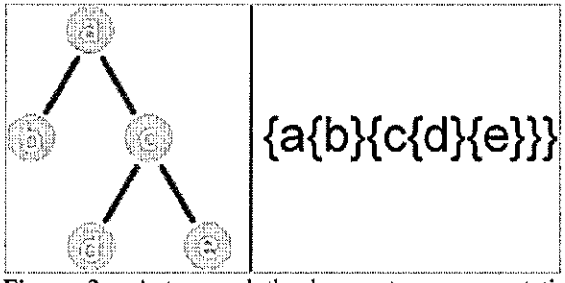


Figure 3. A tree and the braces tree representation belong to this tree.

An ordinary SRR, HTML code, braces tree, and simplified braces tree are as shown in Figure 4.

Figure 4. An ordinary SRR, HTML code, braces tree and simplified braces tree representation

Because of the structure of HTML document, a SRR DOM may be very complex. Therefore, the tree structure should be simplified by a content flattening process as shown in Figure 5.

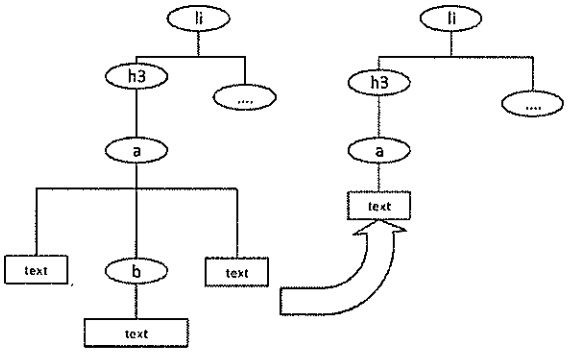


Figure 5. A section from SRR DOM tree and its simplification

The simplification process flattens and merges all content based sub trees to the upper level node in order to simplify similarity comparison.

In some cases, some irrelevant (i.e. page numbers) information has been rendered in the same node with irregular results. To eliminate non-SRR information, irregular results of chosen node are evaluated by extra filters.

4 EXPERIMENTAL RESULTS

The proposed method has been evaluated on leading search engines with a keyword list. The keyword list has been built manually from different domains. The number of used keywords is 500 and domains are listed below:

- Google Search Trends (150 keyword)
- AOL Search Trends (100 keywords)
- Indeed.com (10 trending job keywords)
- Amazon.com (top seller 15 items)
- Ebay (top 20 search keywords)
- Twitter (top trends 5)
- IMDB (top 200 movie title)

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Experimental results show that 88 percent of the search results include at least one irregular SRR.

Three different search engines have been used for testing the proposed method. Precision and recall have been calculated for all search engines. Precision and recall equation is shown below:

$$precision = \frac{|true\ positives|}{|true\ positives| + |false\ positives|} \quad (2)$$

$$recall = \frac{|true\ positives|}{|true\ positives| + |false\ negatives|} \quad (3)$$

Table 3. Precision and recall values for different search engines

	precision	recall
Search engine 1	%98.3	%97.1
Search engine 2	%97.5	%96.4
Search engine 3	%90.6	%87.8

According to experimental results, a well-designed search result page has been obtained from the proposed method.

5 CONCLUSIONS

A new method has been developed for finding and extracting the SRRs. The method first detects content dense nodes on HTML DOM and then extracts SRRs.

Experimental results show that the proposed method is successful for finding and extracting the SRRs without any human interaction and manual process.

In future works, with the help of semantic concepts, the result may align to a type automatically.

ACKNOWLEDGEMENT

This research was supported by the Huawei Technologies Co., Ltd. and Ministry of Science, Industry and Technology (Project Number 0431.STZ.2013-2).

REFERENCES

[1] J. Madhavan, S.R. Jeffery, S. Cohen, X.L. Dong, D. Ko, C. Yu, A. Halevy, "Web-Scale Data Integration: You Can Only Afford to Pay As You Go", Proc. Conf. Innovative Data Systems Research (CIDR), pp. 342-350, 2007.

[2] H. Zhao, W. Meng, Z. Wu, V. Raghavan, C. Yu, "Fully Automatic Wrapper Generation for Search Engines", Proceedings of the 14th International Conference on World Wide Web, pp.66-75, 2005.

[3] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRunner: Towards Automatic Data Extraction from Large Web Sites", VLDB Conference, pp.109-118, 2001.

[4] W. Liu, X. Meng, W. Meng, "Vide: A Vision-Based Approach for Deep Web Data Extraction", IEEE Transactions on Knowledge and Data Engineering, 22 (3), pp.447-460, 2010.

[5] E. Ferrara, P. De Meo, G. Fiumara, R. Baumgartner, "Web Data Extraction, Applications and Techniques: A Survey", arXiv preprint arXiv:1207.0246, 2012.

[6] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, A. Halevy, Google's Deep Web Crawl, Proceedings of the VLDB Endowment, 1(2), pp.1241-1252, 2008.

[7] A. Banu, M. Chitra, "DWDE-IR: An Efficient Deep Web Data Extraction for Information Retrieval on Web Mining", Journal of Emerging Technologies in Web Intelligence, 6(1), pp.133-141, 2014.

[8] G. Miao, J. Tatemura, W.P. Hsiung, A. Sawires, L.E. Moser, "Extracting Data Records From the Web Using Tag Path Clustering", In Proceedings of the 18th International Conference on World Wide Web", pp.981-990, 2009.

[9] R.B. Trieschnigg, K.T.T.E Tjin-Kam-Jet, D. Hiemstra, "Ranking XPaths for Extracting Search Result Records", Technical Report TR-CTIT-12-08, Centre for Telematics and Information Technology, University of Twente, Enschede. ISSN 1381-3625, 2012.

[10] X. Yin, W.Tan, X. Li, X., Y.C. Tu, "Automatic Extraction of Clickable Structured Web Contents for Name Entity Queries", In Proceedings of the 19th International Conference on World Wide Web, pp.991-1000, ACM, 2010.

[11] M.K. Bergman, "The deep Web: Surfacing Hidden Value", BrightPlanet.com [Online].

[12] A.H. Laender, B.A. Ribeiro-Neto, A.S. da Silva, J.S. Teixeira, "A Brief Survey of Web Data Extraction Tools", ACM Sigmod Record, 31(2), pp.84-93, 2002.

[13] C.H. Chang, M. Kayed, M.R. Girgis, K.F. Shaalan, "A Survey of Web Information Extraction Systems", IEEE Transactions on Knowledge and Data Engineering, 18(10), pp.1411-1428, 2006.

[14] E. Ferrara, P. De Meo, G. Fiumara, R. Baumgartner, R., "Web Data Extraction, Applications and Techniques: A Survey", arXiv preprint arXiv: 1207.0246, 2012.

[15] A. Sahuguet, F. Azavant, "Building Intelligent Web Applications Using Lightweight Wrappers", Data & Knowledge Engineering, 36(3), pp.283-316, 2001.

[16] L. Liu, C. Pu, W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources", In Data Engineering, Proceedings 16th International Conference, pp. 611-621, 2000.

[17] N. Kushmerick, Wrapper Induction: Efficiency and Expressiveness, Artificial Intelligence, 118(1), pp.15-68, 2000.

[18] I. Muslea, S. Minton, C.A. Knoblock, "Hierarchical Wrapper Induction for Semistructured Information Sources" Autonomous Agents and Multi-Agent Systems, 4(1-2), pp.93-114, 2001.

[19] C.H. Chang, S. C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery", In Proceedings of the 10th International Conference on World Wide Web, pp. 681-688, AC, 2001.

[20] B. Liu, R. Grossman, Y. Zhai, "Mining Data Records in Web Pages", In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.601-606, ACM, 2003.

[21] A. Arasu, H. Garcia-Molina, "Extracting Structured Data From Web Pages", In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, pp.337-348, ACM, 2003.

[22] Y. Zhai, B. Liu, "Web Data Extraction Based On Partial Tree Alignment", In Proceedings of the 14th International Conference on World Wide Web, pp.76-85, ACM, 2005.

[23] M. Pawlik, N. Augsten, "A Memory-Efficient Tree Edit Distance Algorithm", In Database and Expert Systems Applications, pp.196-210, Springer International Publishing, 2014.

[24] E. Estrada, N. Hatano, M. Benzi, The physics of communicability in complex networks, Physics Reports, 2012, vol. 514, 89—119

[25] Y. Shang, Perturbation results for the Estrada index in weighted networks, Journal of Physics A: Mathematical and Theoretical, 2011, vol. 44, no. 7, 075003

[26] E. Estrada, The Structure of Complex Networks: Theory and Applications, 2011, Oxford University Press

Server Monitoring Using Android Devices

Negar Shakeribehbahani, Nor Azlina Abd Rahman, Kamalanathan Shanmugam, Payam Nami
Asia Pacific University of Technology and Innovation
Kuala Lumpur, Malaysia
negarshakeri@gmail.com ; nor_azlina@apu.edu.my ; kamalanathan@apu.edu.my ; paynam@gmail.com

Abstract— **Server Monitoring Using Android Devices is an application which Using Android operating system devices to monitor Windows servers. This software allows network administrators to monitor resources and the status of the servers, such as CPU utilization, RAM usage, Hard Drive I/O activity, storage space and running services and processes remotely and easily through an Android phone or tablet by even a poor internet connection.**

Keywords-component; Server Monitoring; Windows Server Monitoring; Android Application; Andoroid Devices; SOAP Technology

I. INTRODUCTION

Nowadays, World Wide Web (WWW) through the Internet has been changed the aspects of human life, like communication ways, knowledge and cultures. In additional, business owners and investors have appreciated the importance of Internet. Among them, Internet and technologies have been grown excessively. So, that the communication ways are much easier to use. Today, there are lots of server machines which provide the network services such as web servers, file servers, database servers and more. In order to ensure the servers function properly and provide high quality of services, they need to be monitored by network administrators all the time. Besides, most people are using smart phones and tablets with a variety of applications and mobile internet connections, thus server monitoring application via Android would be a useful solution especially for whose positions related to networking. By using this application, network administrators are able to monitor multiple servers anytime and anywhere, and meanwhile, they can spend their time on doing other tasks far from the servers.

A. Rationale

Nowadays remote access to computers via the mobile devices has become more popular. The reasons are that it is more convenient to carry a cell phone rather than computer, and also, employees need to have access to their documents remotely. Accordingly, a server administrator needs to access the server to monitor its resources and status to make an appropriate action in the event that the server is suffering a problem or crash.

In order to minimize the current payments for server monitoring, a solution which makes the server administrators able to monitor the servers (In this case Windows servers) remotely, quickly and easily and consequently, minimizing the number of employees who monitor the server on the company's premises may help. "Server Monitoring Using Android Devices" is software which allows a server administrator to monitor the server and its resource status, such as remaining storage space, memory usage, CPU usage, and hardware device temperatures remotely.

B. Problem Statement

In order to ensure customers' satisfaction and avoiding business losses, servers need to be monitored continuously to minimize the downtimes and maintain the performance. However, hiring persons to monitor the company's servers needs a large amount of budget in the aspect of human resource. Indeed, the company is paying for the employees who have nothing to do in most of the times.

On the other hand, monitoring CPU temperature aids, to avoid damage, low response time and short lifespan of the system. Moreover, monitoring usage of RAM helps avoid low response time. Furthermore, monitoring status of Hard Drive helps to monitor activity and availability of the Hard Drive in order to avoid high activity and low disk space which can be due to decrease the performance of the server, even make it unresponsive. Additionally, monitoring Bandwidth Usage prevents bandwidth bottlenecks and increase the response time of the server.

C. Aim

Server Monitoring Android Application consists of designing and implementing a system to allow network administrators to monitor the server and its resources status such as temperature and utilization of CPU, status of RAM, status of Hard Drive I/O activity, status of bandwidth usage of network devices, server storage and running services and processes remotely. Besides that, the network administrators are able to view the history of the servers as well. This application is able to connect to the server automatically in desired intervals to retrieve and record server information. The application allows the network administrators to monitor:

- Temperature of CPU remotely.
- Utilization of CPU remotely.
- Status of RAM remotely.
- Status of Hard Drive I/O activity remotely.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

- Status of bandwidth usage of network devices remotely.
- Server storage remotely.
- Running services and processes remotely.

II. METHODS

This project consists of two major steps: research and development.

A. Research

The Stage of the research is collecting knowledge and information from journals, conference papers, books and other valid academic resources. Also, questionnaires via internet or hard copy user surveys are used to collect information. Based on the distributed questionnaire to target respondents who were 25 persons including network administrators, network managers and IT managers, all the preliminary requirements have been gathered and analysis of questionnaire aids collected information to be meaningful. Analysis answers were useful to find the requirements for this proposed system. Through questionnaire analysis some items such as popularity of Android and Windows operating system in network industry, reasons of server monitoring, important metrics which application should support, and finally the willingness of network administrators are clarified.

B. Evaluating a Popular Monitoring System

OpManager network monitoring which is web-bases software is introduced. OpManager monitor network in several aspects which covers fault, Configuration, Accounting, Performance, Security (FCAPS) including performance of network and server monitoring, fault management, security management, configuration management and accounting management. OpManager monitors the Active Directories on the Windows Servers and shows the statistics of them about the number of users, computers, inactive users, active users domain based and aggregately up on monitor request. It is also capable to collect the information from Unix-based operating systems and their accounting status. OpManager monitors the event logs of the servers constantly. By monitoring this logs, especially the security logs of the servers, it alerts the monitors in the case that a suspicious activity detected. Moreover, it monitors the firewalls including hardware and software, and also antivirus software all over the network. For example, it alerts the monitors if antivirus software is out of date on a specific computer or a firewall is not configured properly. Furthermore, it is capable to monitor specific files or folders on any server or computer for security reasons. It alerts the monitors in the case of suspicious activities.

OpManager will be installed on a Windows server as a Windows service. It enjoys a web-based graphical user interface by bundled web server by the software installer. The TCP/IP port used to serve the GUI is selectable by the system administrator during installation. The main advantage of a web-based GUI is that the OpManager interface will be accessible across the network by the monitor employees. Furthermore, OpManager may have more than one user [1].

C. Comparison of OpManager and Server Monitoring Using Android Devices

Table 1 shows the comparison of OpManager and Server Monitoring Using Android Devices.

Table 1. Comparison of OpManager and Server Monitoring Using Android Devices

Software Specification	OpManager	Android Application
Platform	Windows, Linux	Android
UI	Web	Android application
Installation	Needs server	Does not need server
Portability	Medium	High
Protocol	SNMP	SOAP
Devices	Router, Server, Switch	Server
Cost	Minimum 1995 $	About 10 $

As the table 1 shows these two systems are quite different from each other. They have the same goal which is monitoring, but their performances are different. The platform used for OpManager is Windows or Linux and it needs to be a dedicated server while in this system is Android and it does not need specified server. Also, user interface in OpManager is Web, and it uses SNMP protocol, but in this system would be Android application and the protocol is SOAP. Moreover, OpManager is able to monitor routers, servers, and switches while this proposed system can monitor only servers. On the other hand, the portability of OpManger is medium contrast this system is high. Finally, the price of OpManager compares to this system is much higher. Therefore, even though OpManager offers more features, this system is designed to be personal, simple, easy to use and portable. Also, in the aspect of prerequisites, the proposed system does not require high cost hardware and can be installed and used by anyone with a low networking knowledge.

D. Prototype Development

Server monitoring using Android device prototype is distributed software which consists of client-side application known as android application, and server side application which is a web application.

Server side application provides the information regarding the status of a server such as CPU usage, memory usage, network load etc. for android application installed on network

ISBN: 978-1-941968-02-4 ©2014 SDIWC

administrator's smart device. The technology used for interoperability is Simple Object Access protocol (SOAP).

Figure 1 illustrates the Use Case diagram for Android application prototype. This prototype is designed specifically for the network administrators to monitor the servers remotely using Android devices. This application enable the network administrators to handle their tasks effectively and more convenient as monitoring and decision making can be done at anytime and anywhere.

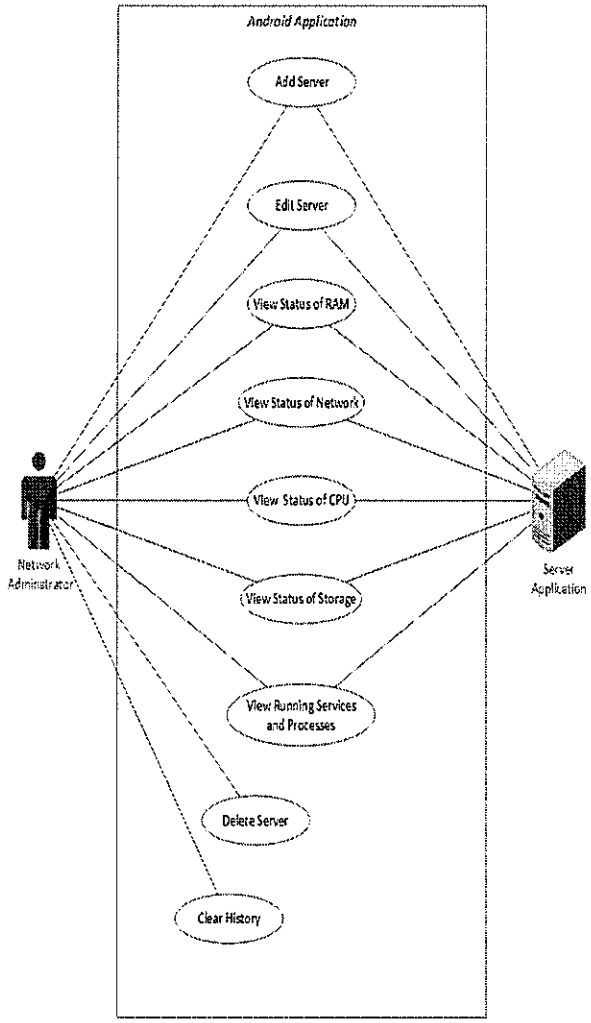


Figure 1. Use Case for Android Application

Figure 2 illustrates the Use Case diagram for server application of this project.

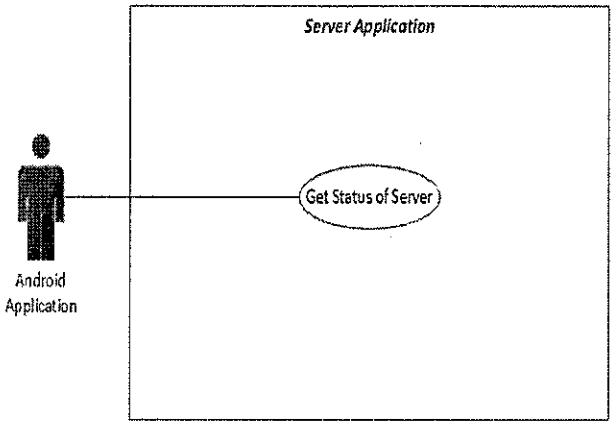


Figure 2 - Use Case for Server Application

III. TECHNOLOGIES USED FOR PROTOTYPE DEVELOPMENT

The prototype developed using several integration of technologies as discuss in this section.

A. Programming language

Server Monitoring Android App project requires two sides of programing; the first part is a Server Side Application and the second part is Client Side (Android) Application. So, each of them requires a different type of programming language. For Server Side Application C# to develop a web service and For Client Side Application Java would be suitable programming languages to an Android application.

B. Android Development

Android is a new open source platform for mobile devices which produced by Google company. Android is designed to be a complete software stack. It includes an operating system, middleware, and core applications. It is designed to facilitate the development process. Security developers can easily work with and rely on flexible security controls. Also, it is a very intuitive operating system, and users are able to understand easily how applications work and control applications.

C. Java Programming Language

All Android applications are natively developed by using Java programing language. Java has powerful libraries which they are created to help developer to build applications easier. Java enjoys several advantages. Firstly, it is object oriented. Secondly, it is easy to learn, understand and use. Moreover, it is developed and designed to be an independent platform, and to be secure and use a virtual machine to run on multi-platform. Finally, it has special strength in the servers and middleware. Typically, Android applications are developed by using Java and benefit these advantages. For example Android SDK has a large number of standard Java libraries such as networking libraries, graphics libraries, data structure libraries and more. So, it helps developers to build awesome Android applications [2].

D. Using Android Virtual Device for Android Development

AVD (Android Virtual Device) is a device configuration and allow the developers to model actual device by recognizing hardware and software. They are able to emulate by Android Emulator. Moreover, AVDs able to be configured and run on any version of Android. Therefore, AVD let developer to create several configurations to test many hardware and Android platforms. When the application is running on the emulator, it will able to use the services of Android platform such as play audio and video, store and retrieve data, access the network and create themes. Also, developer can use a variety of commands and options to control and change the performance of the application [3].

E. C# Programming Language

C# (pronounces as "C sharp") is one of the most powerful programming languages, recent nine years created by Microsoft for the major part of its .NET initiative. C# developers created it with inspiration from Java and C++ programming languages. It is multi paradigm, object oriented programming language and suitable for deployment in a distributed environment. Also, it helps developer to create portable applications. It can support the principles of software engineering such as array bounds checking, strong types and automatic garbage collection. C# is able to use XML (Extensible Markup Language) and SOAP (Simple Object Access Protocol) to allow access to programming object, so the programmer does not need to write extra code. Using C# programming language makes the process of developing faster and less expensive [4].

F. Web Services In .Net

Web services are methods for communication between devices. Web services help developer to create distributed application much easier to design and develop. In fact, web services enable a remote consumer to run it based on the given parameters and get the result over HTTP protocols. It can be considered as a set of remote functions for a developer. [5] The advantage of web services is that they are not relative to specific technology because they use XML or other global common independent languages for communication. On the other hand, developer can use it in most of the development scenario. [6] .NET uses the web services as the main protocol to establish communication between applications [7].

G. SOAP Web Services

SOAP is a lightweight protocol for transmitting information in a decentralized, distributed environment. SOAP is an Extensible Markup Language (XML) based protocol which contains three parts: an envelope that defines a framework for describing what is in a message and how to process it, a set of encoding rules for expressing instances of application-defined data types, and a convention for representing remote procedure calls and responses [8].

In aspect of security, SOAP benefits a high level of security. Authentication and authorization in sending messages can be applied to SOAP web services. Furthermore, encryption can be applied to the messages between the source and destination to secure the message, for example using the SOAP over the HTTPS protocol instead of HTTP can encrypt the message to ensure that the message cannot be modified and read the middle of the way. [9]

Figure 7. SOAP Web Services

H. Advantages of SOAP over DCOM

There are different protocols proposed for communication among software components of the distributed software. Distributed Component Object Model (DCOM) and Simple Object Access Protocol (SOAP) are popular protocols which are mainly developed by Microsoft. One popular protocol is DCOM. DCOM is a major methodology in distributed computing on the Windows platform. Although it makes developers' work less difficult by hiding many complexities of client-server application development, DCOM has two major disadvantages. Firstly, it is only mature on Windows and is not suitable for cross platform communications. Second, implementing DCOM applications in a corporate environment are difficult where communication needs to be performed across firewalls. In contrast, another technology which is SOAP is based on two protocols: XML and HyperText Transfer Protocol (HTTP), a variety of platforms, such as Windows and Android are compatible with these protocols [10].

In summarize, SOAP is compatible with all platforms and it functions on the HTTP protocol, it is a more appropriate option for interoperability.

I. SQLite Database Management System

SQLite database does not need a setup procedure or administration of the database in Android. Android is able to support SQLite database completely. After defining the SQL statement to create and update the database, the Android platform will manage database automatically. The database will be saved in the application data directory by default, if an application creates a new database. It is an open source database and able to support relative database features such as, SQL syntax, transaction and prepared statements. SQLite does not require amount of large memory at runtime, so it's a good choice for embedded database. SQLite provides type of data like, TEXT, INTEGER and REAL which they are comparable to STRING, LONG and DOUBLE in Java.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

IV. PROTOTYPE

This section demonstrates the prototype of the servers monitoring android application and describes the feature implemented visually.

Add button is to enable the network administrators to add one or more servers in the list of server of application. Network administrators are able to monitor one or several numbers of servers at the same time. After the connection is successful, the application will retrieve all information that needed such as the name of the server, detail of devices, current status of RAM, CPU and CPU temperature and more. Figure 3 and Figure 4 shows add server screenshot of the prototype.

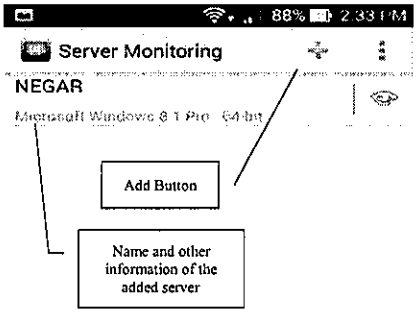


Figure 3. Add Server

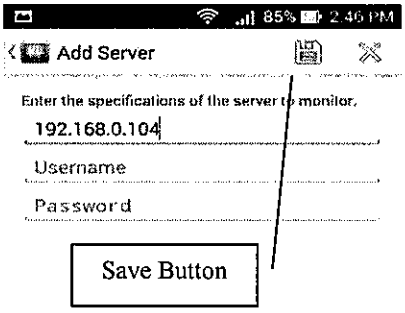


Figure 4. Add Server

This prototype enables network administrator to update the previous added information of the servers. The edit button is used to edit the information such as IP address, Username and password. Figure 5 shows the edit server screenshot of the application.

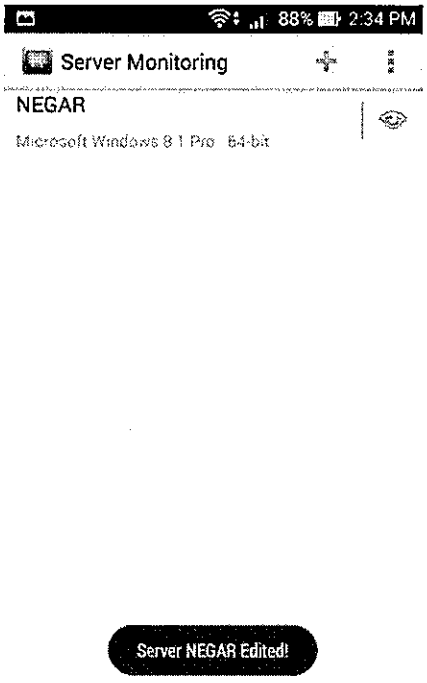


Figure 5. Edit Server

The prototype able to retrieve current status of RAM and virtual memory of the server and displays this information in chart and graph form to make the monitoring process become easier and clearer. Moreover, this application is capable to show date, time and history of RAM in graph form. It can illustrate the last hundred screenshots of RAM information.

Refresh button used to refresh current page and view the real time status of the servers' information retrieved. Furthermore, this prototype cans automatically connecting to the server in desired intervals to update the information. If the Internet has a problem and network administrator could not connect to server to get current situation of the server, the information of the servers can be retrieved from the servers' history. Figure 6 shows the screenshot to view the current status of RAM and the history of the servers.

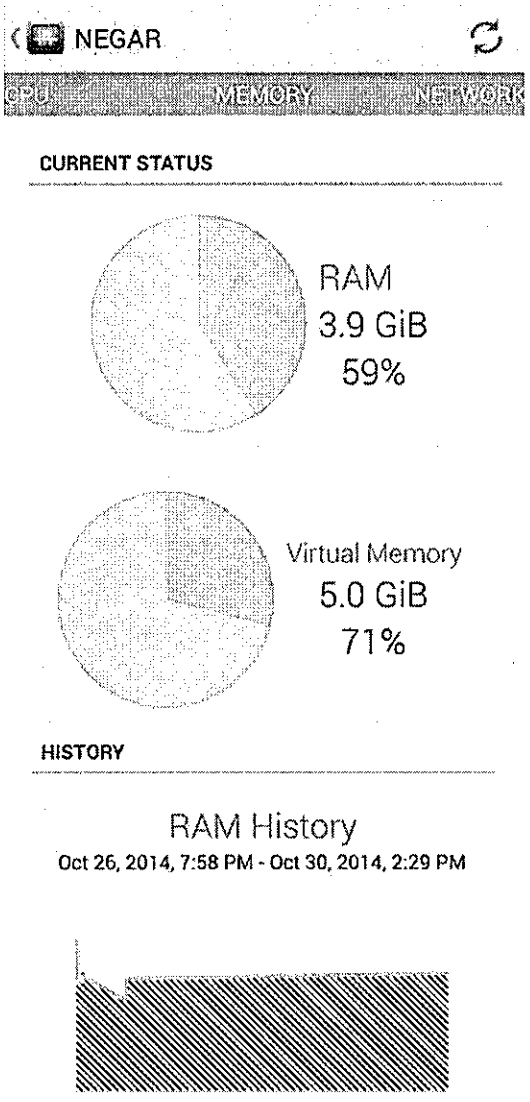


Figure 6. View Current Status of RAM and History

Another functionality of this prototype is to retrieve current status of bandwidth usage of network devices such as Ethernet upload and download, and Wi-Fi upload and download of the server and specific speed number (Mbps) and displays the information in a diagram form to the network administrator. Figure 7 and Figure 8 show the screenshots to view the current status of network and the history of the Ethernet.

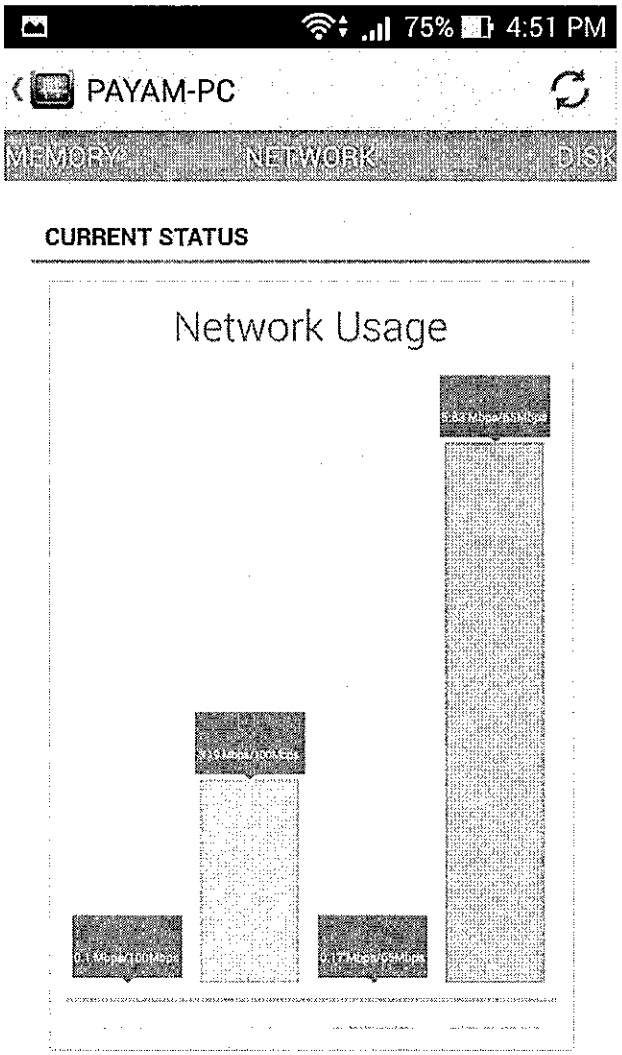


Figure 7. View Current Status of Network and History (Ethernet)

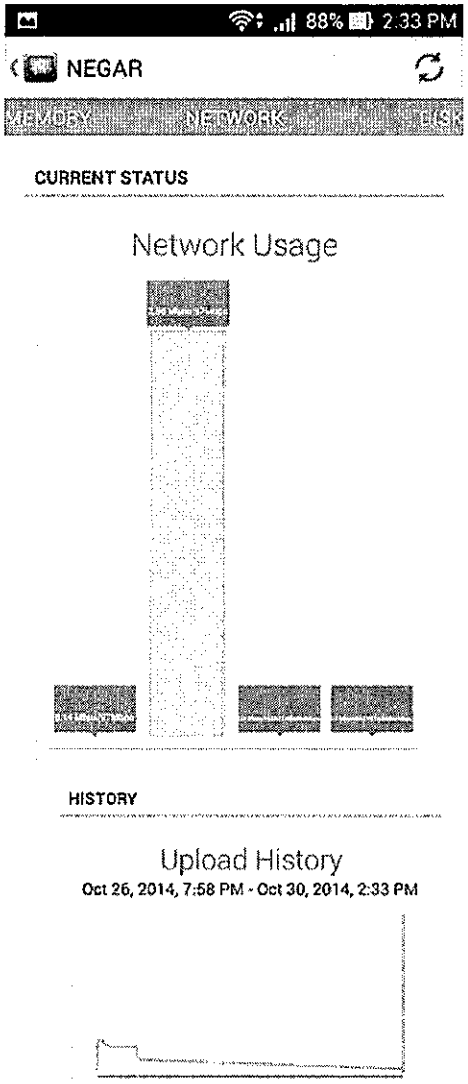


Figure 8. View Current Status of Network and History (Wi-Fi)

Next functionality of the prototype is to retrieve current usage of CPU and CPU temperature of the server and displays the information to network administrator. Figure 9 shows a screenshot to view the current status of CPU and the history.

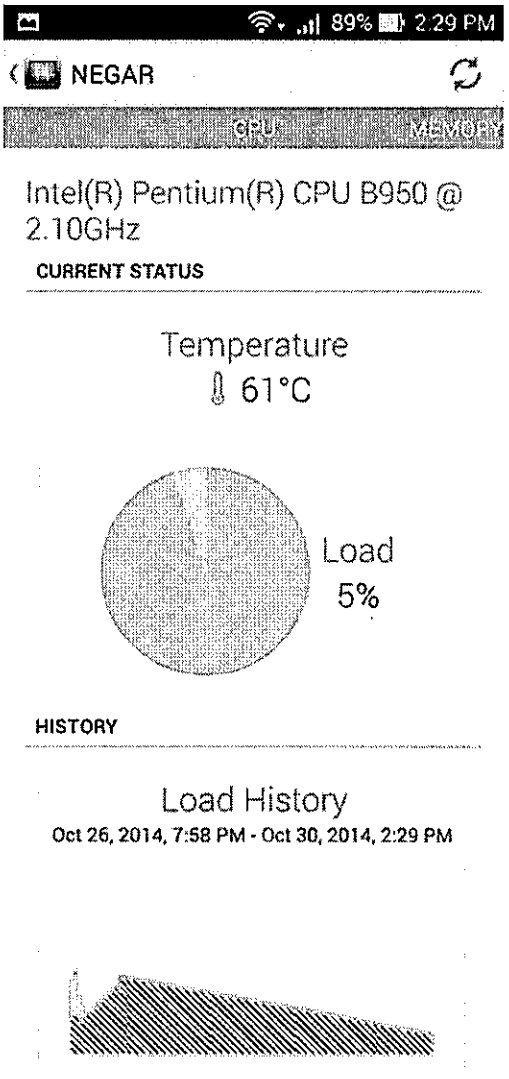


Figure 9. View Current Status of CPU and History

Current usage of storage and hard drive I/O activity of the server can be retrieved using this prototype. Figure 10 shows a screenshot to view the current status of storage and the history.

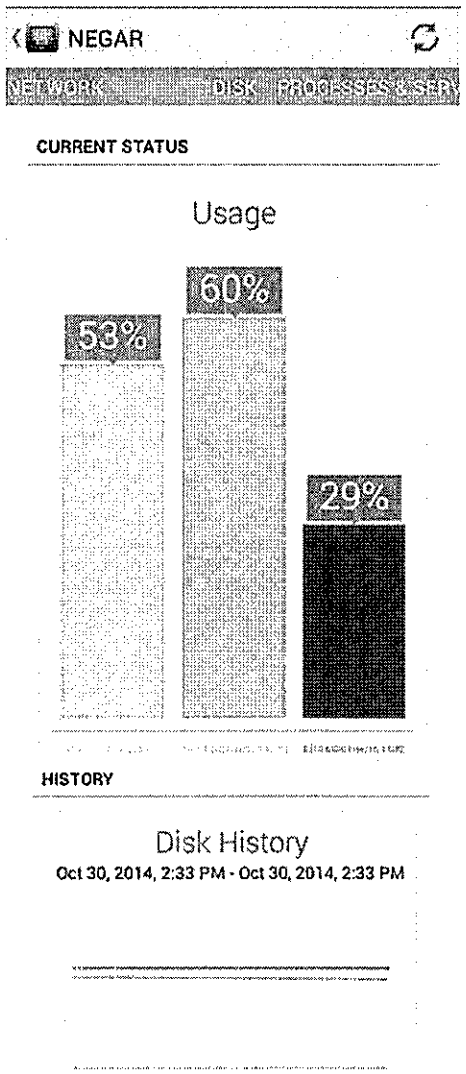


Figure 10. Views Current Status of Storage and History

Figure 11 shows a screenshot to view current status of running services and processes of the servers. This is one of the functionalities of this prototype that enable to retrieve current status of running services and processes of the server and displays the information to the system administrator, so that any action can be taken if there is any inappropriate information showed.

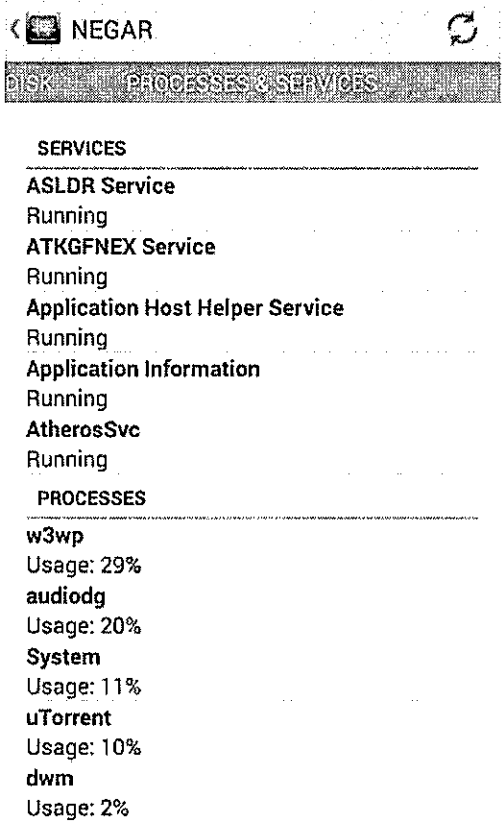


Figure 11. View Current Status of Running Services and Processes and History

This prototype also provides functionality that enables the network administrator to delete the previous added server from the list of servers. To delete the server user should select the delete button and system will delete the chosen server with all the information from the list of servers. Figure 12 shows the menu for delete, edit and clear history screenshot of the prototype.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

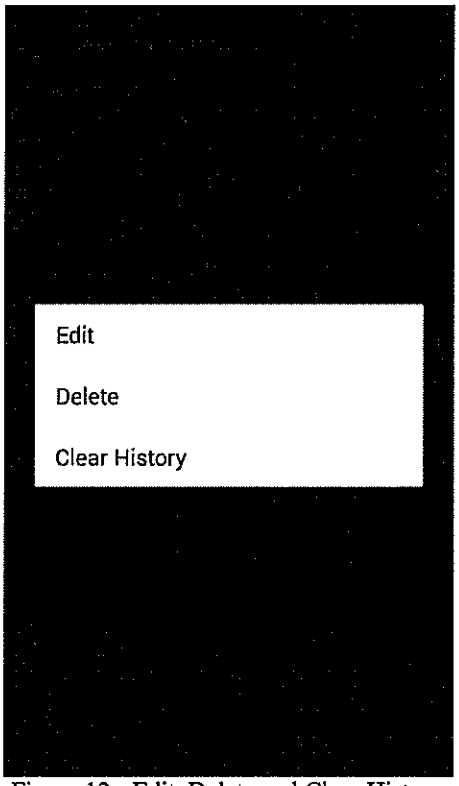


Figure 12 - Edit, Delete and Clear History

Figure 13 shows that server has been deleted from the list of severs.

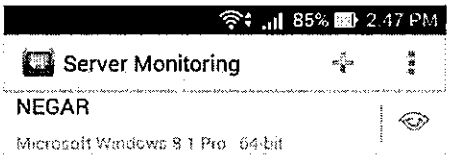


Figure 13 - Delete Server

Clear the history of the server is another functionality of the prototype. Clear history button is used to clear the history of the servers that consist of all information about CPU, RAM, network, storage, running services and processes of the chosen server. Figure 14 shows that history of the chosen server has been cleared.

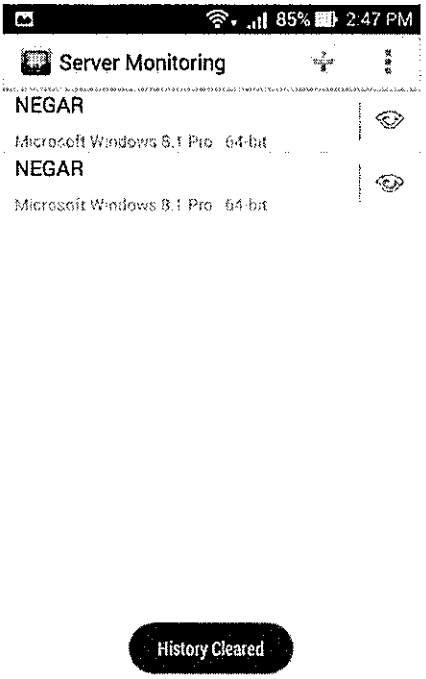


Figure 14 - Clear History

This prototype is able to show all the needed information of the servers to network administrators but it is requires establish the communication between server side application and client side application. As mentioned before developer used SOAP to create connection between server and client to get the information from the server and then displays the information to network administrators. Figure 15 shows application is trying to connect to the server to retrieve the information of the servers connected

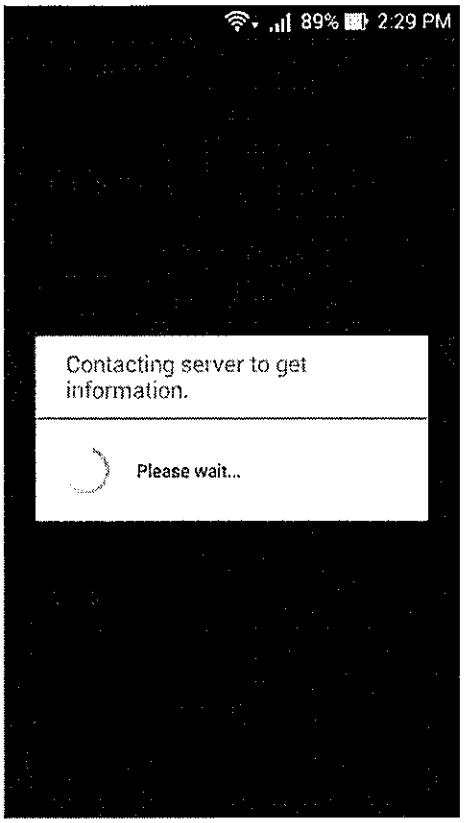


Figure 15 - Server and Client Interoperability

The auto update service feature enables the application to collect the status information of registered servers in the application as a background task. In other words, this service collects the server information while even the application is not running in settable intervals. It helps to have an updated data of servers to illustrate the more meaningful history for the servers.

All the functionalities of this prototype as shown in figure 5, figure 6, figure 7, figure 8, figure 9, figure 10 and figure 11 enable the network administrator to monitor the status of the server and display their history in graph form. In fact, the graphs can illustrate the last hundred screenshots of the server status. Besides that, all the server specifications are also provided with a Refresh button to retrieve the latest status of the server and view the real time status of the servers' information retrieved. Furthermore, this software application can automatically connect to the server in desired intervals to update the information. If the Internet has a problem and network administrator could not connect to server to get current situation of the server, the information of the servers can be retrieved from the servers' history.

V. FURTHER FEATURES

Considering of this prototype improvement, several features such as Sending smart alerts, through SMS or E-mail by recognizing potential failures and support SMTP protocol beside SOAP are recommended.

VI. CONCLUSION

The main idea behind this system is monitored devices on Windows servers via Android application. This prototype permits network administrators to monitor resources and the status of the servers, such as CPU utilization, RAM usage, Hard Drive I/O activity, storage space and running services and processes remotely and easily through an Android devices. This system would be a beneficial application for those who are working in the networking field. They are able to monitor multiple servers anytime and anywhere, and meanwhile, they can spend their time on doing other tasks far from the servers. Moreover, it helps to reduce the downtime of servers and network, and it can be due to increasing level of customers' and employees' satisfaction.

ACKNOWLEDGMENT

The authors thank the Head of Computing & Technology, Dr Thomas Patrick O'Daniel for his constructive comments and suggestions that were vital in improving the quality of our paper. The authors also wish to express gratitude to the management of Asia Pacific University for their support.

REFERENCES

[1] Zoho Corporation. (2014). *ManageEngine: Enterprise IT Management | Network Management Software*. Retrieved January 02, 2014, from http://www.manageengine.com/

[2] Burd, B. (2014). *Java Programming for Android Developers* (First ed.). (B. Buikema, Ed.) New Jersey: John Wiley & Sons, Inc.

[3] Felker, M. B. (2012). *Android Application Development* (Second ed.). (R. Senninger, Ed.) New Jersey: John Wiley & Sons, Inc.

[4] Troelsen, A. (2012). *Pro C# 5.0 and the .NET 4.5 Framework* (Sixth ed.). (E. Buckingham, Ed.) New York: Paul Manning.

[5] Fraser, S. R. (2009). *Pro Visual C++/CLI and the .NET 3.5 Platform* (First ed.). (M. Moodie, Ed.) New York: Apress.G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[6] MacDonald, A. F. (2003). *Programming .NET Web Services* (First ed.). (N. K. Osborn, Ed.) California: O'Reilly Media.

[7] Jens Bertram, C. K. (2012). Secure Web Service Clients on Mobile Devices. *Procedia Computer Science, 10*, 696-704.

[8] Box, D., Ehnebuske, D., Kakivaya, G., Layman, A., & Mendelsohn, N. (2000). Simple Object Access Protocol (SOAP) 1.1. San Francisco: W3C.

[9] Waleed, G., & Ahmad, R. (2008). Security protection using simple object access protocol (SOAP) messages techniques. Penang: Electronic Design, 2008. ICED 2008.

[10] Davis, A., & Zhang, D. (2002). A comparative study of DCOM and SOAP. Multimedia Software Engineering.

An Approach to Detect Spam Emails by Using Majority Voting

Roohi Hussain
Department of Computer Engineering,
National University of Science and Technology,
H-12 Islamabad, Pakistan

Usman Qamar
Faculty, Department of Computer Engineering
National University of Science and Technology,
H-12 Islamabad, Pakistan

ABSTRACT

Internet usage has become intensive during the last few decades; this has given rise to the use of email which is one of the fastest yet cheap modes of communication. The growing demand of email communication has given rise to the spam email which is also known as unsolicited mails. In this paper we propose an ensemble model that uses majority voting on top of several classifiers to detect spam. The classification algorithms used for this purpose are Naïve Bayesian, Support Vector Machines, Random Forest, Decision Stump and k-Nearest Neighbor. Majority voting generates the final decision of the ensemble by obtaining major votes from the classifiers. The sample dataset used for this task is taken from UCI and the tool Rapidminer is used for the validation of the results.

KEYWORDS

Spam email, filtering, Naïve Bayesian, SVM, Random Forest, Decision tree, Rapidminer

1 INTRODUCTION

Internet usage has become intensive during the last few decades; this has given rise to the use of email which is one of the fastest yet cheap modes of communication. However the rise of email and internet users resulted in the striking increase of unsolicited bulk/spam emails. Spam emails are the junk emails that are sent to numerous undisclosed recipients and that contains identical messages for everyone. Botnet, which is group of programs communicating with other similar programs, is specifically used to send spam emails and it is known for its malicious implication.

The enormous amount of spam data effects the Information Technology based businesses and brings loss of billions of Dollars to the organizations in terms of its output [1]. In last few years, spam emails have become a source for intruding the sensitive data and this posed a serious threat to the sanctuary of many departments [2].

Researchers used classification that focuses on three levels of the email i.e. email address, subject line and body contents. Content based spam detection is the most effective of all three.

The aim of this paper is to propose an ensemble that uses majority voting approach in combination with filtering algorithms for spam detection.

1.1 Spam Features

Spam emails have following features [3], the emails are sent to undisclosed recipients for the advertisement of services/products/offensive material. The aim is to deceive innocent people by gaining personal data of the masses and abuse it. Majority of the spam emails do not offer unsubscribe option.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

1.2 Type of Spam Filters

A SPAM email filter consists of instructions which can block the emails; however the blocking depends on the type of data the email holds. Spam filter like any other filter has the task to classify emails as spam or ham. For more accurate results the sets of instructions in the filter needs to be precise which can determine clearly what can get through. The spam filters are of many types like address filter, subject line filter, content filter or word filter.

Outline of this paper

In Section 2, related work of different researchers for the email spam detection is discussed, Section 3 the proposed model is discussed in detail, Section 4 details the classifiers used in the proposed methodology, Section 5 is about results obtained by testing the sample dataset on the proposed model and finally the last section discusses the conclusion and future work.

2 RELATED WORK

Goodman et al. [4] gave a view of what anti-spam protection is. A brief history was given about spam and main areas of research in this field were discussed. They were of the view that learning based algorithms used with hoaxing and tricky technologies can be used to overcome this problem in near future.

Siponen and Stuckye [5] defined different approaches used for controlling spam. They came to the conclusion that filtering is the most effective way of controlling spam. Wang and Cloete [6] debated over email classification along with spam filtering and discussed the filtering techniques in more detail.

In this paper [7] the researchers applied Naïve Bayes email classification techniques successfully. Other Bayesian classifiers have also been applied to filter spam [8]. In this paper [9] the researcher proposed techniques to improve the performance of Naïve Bayes classifier.

Another approach used by researcher is the memory based anti-spam filtering. [10] In this technique the training data is stored and then incoming instances are evaluated against the stored data. These techniques are complex and difficult to manage due to the effort required for processing large amount of training data at the time of testing. [11]

Methods such as boosting have shown some outstanding results for spam classification. [12] Boosting is very effective when used with decision trees, although this technique is expensive due to the complexity of base learners but better results are achieved as precision is high.

In this paper [12], weighted voting method along with clustering is used for the detection spam. The sample dataset is first compared with all cluster centroids and then weighted voting technique generates the final decision of the ensemble by obtaining majority votes from the classifiers.

Paper [12] discusses how clustering can be used through criterion function for the detection of spam. Criterion function checks the similarity between email messages in the clusters by using k-Nearest Neighbor algorithm. A new genetic algorithm is proposed for this purpose.

Paper [4] elucidates a novel approach called Symbiotic Data Mining that uses content based filtering with collaborative filtering for spam detection. In this paper [14] spam is classified through Linear Discriminant Analysis where bag-of words document is generated for every incoming website.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

3 FRAMEWORK OF THE PROPOSED MODEL

The overall architecture of the proposed model is given in Figure 1. The steps of the proposed model are discussed briefly in the following section.

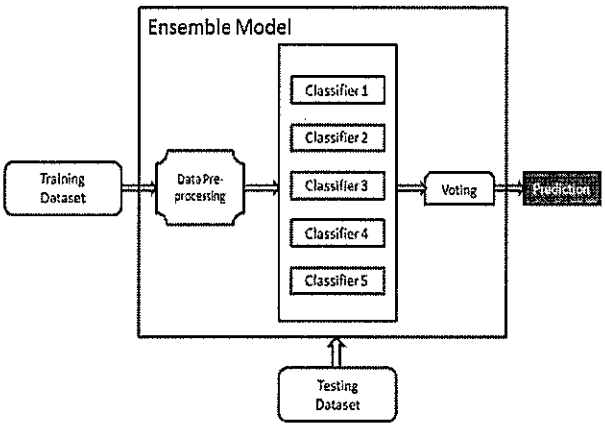


Figure 1: Proposed Ensemble Model

3.1 Dataset

In our approach the spam training data set has been used for the testing purpose, which is taken from UCI website, created by Mark Hopkins, Erik Reeber, George Forman and Jaap Suermondt. There are total 58 attributes (57 continuous attributes, 1 class label) with 4601 instances.The instances have been labeled as spam(1) and non-spam(0) in the label class. Out of total instances of 4601, 39.4% are spams and other is non-spam.

3.2 Data Pre-processing

It is important to perform the data pre-processing step in order to remove any noisy or missing value. Pre-processing task is very important before performing any mining procedure. In our model, we performed two important steps

- Data cleaning
- Normalization

We have not used any reduction techniques in this approach, so the testing is performed on complete data set comprising of 58 attributes.

Training and testing data has been split in 20:80 ratio for testing purpose.

3.3 Study on Classification Algorithms

The base classifiers used for major voting ensemble are as following:

- Naïve Bayesian

This classifier is based on the statistical technique of calculating the probability of an event occurring in the future that can be spotted from the prior occurrence of the similar event. [15] It is a popular algorithm for the classification of the email as ham or spam. Probabilities of word occurrence are the key rule in this algorithm. If some identified words occur more than often in an email; then supposedly it is a spam email. In this technique the filter is trained by using the training data set.

- Support Vector Machine

SVM is based on the theory of decision planes that clearly defines the boundaries for decision making. Pure SVM classifies the data by predicting the class for set of input data. SVM are the supervised learning models what makes use of the learning algorithms to analyze data and identify the patterns. SVM is also known as non-probabilistic linear classifier as it assigns the incoming data to one or more classes. Typical SVMs are only application for data falling under two classes. Therefore, in the case of multi-classes, many binary algorithms have to be applied for the reduction of classes.

ISBN: 978-1-941968-02-4 ©2014 SDIWC

- Decision Stump

The Decision Stump is used for generating a decision tree with only one single split. The resulting tree can be used for classifying unseen examples. The leaf nodes of a decision tree contain the class name whereas a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute.

- K-Nearest Neighbor

This is an example based classifier that uses the training material for comparison. Whenever a new message needs to be classified, the k most similar messages are found in the neighborhood and compared with. If majority of the surrounding messages belongs to one particular class, then the new message is also assigned to that class otherwise not.

The nearest neighbors are found by using the indexing techniques which saves the time for comparison. This classifier is also termed as memory-based classifier as all the training material is stored in the memory [16]. This is a memory based classifier in which complexity is increased when training dataset increases.

- Random Forest

Random forest is a learning algorithm that is used for classification and regression of data. It is formed by the combination of many decision trees at the training phase. Random forest incorporates the bagging method and random selection of features that is used to make decision trees with controlled variance. [17] Prediction through RF is achieved by considering the voted classes of all decision trees and the class with majority votes weighs the output of prediction. [18]

3.4 Majority Voting

The majority voting is also known as average voting which is based on the predictions of the inner base classifiers. The class uses n+1 base classifiers and generates n different models. Each model predicts the outcome of the testing instances. The predictions of all models are taken into account and voting is done to predict the final outcome of the particular instance. If for 4 out of 5 classifiers predict an email as spam, then output labels the email as spam as it received major votes.

4 TESTING RESULTS AND PERFORMANCE ANALYSIS

This section focuses on the testing of the proposed ensemble model. Testing of the proposed model has been performed on Rapidminer software tool.

4.1 Steps for Testing

Step 1: Ten classifiers chosen randomly and tested on the data set to check their accuracy, classification error, precision, recall and execution time. Training to testing ratio used is 10:90. Results shown in table 1.

Table 1. Classifier test results

Category		Spambase dataset with 58 attributes & 4601 instances			
	Precision %	Recall %	Accuracy %	Classification Error %	Time
LDA	51.9	51.81	55.5	44.49	34 s
CHAID	30.29	50.01	60.58	39.42	20 s
ID3	71.26	68.85	63.87	36.13	27 s
Random	78.15	66.01	72.44	27.56	23 s
DT	83.24	70.10	76.09	23.91	20 s
DS	83.4	70.2	76.23	23.77	20 s
k-NN	78.79	80.13	78.99	21.01	24 s
RF	84.11	79.73	82.73	17.27	22 s
SVM	87.74	82.09	85.24	14.76	23 s
Naïve	95.32	94.94	95.71	4.29	26 s

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Step 2: Five classifiers with minimum classification error and highest accuracy are picked up to be used in the ensemble model. Figure 2 shows the graphical result of the classifier accuracy

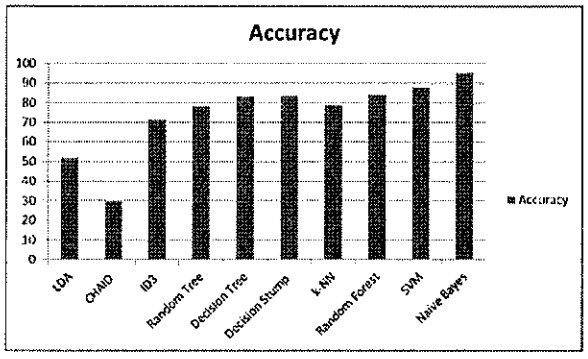


Figure 2: Classifier Accuracy

For testing of the ensemble model, five classifiers have been chosen based on the highest accuracy, precision, recall and processing time.

The five classifiers therefore are;

- Decision Stump
- k-Nearest Neighbor
- Random Forest
- Support Vector Machine
- Naïve Bayes

CHAID yielded less processing time but it was not selected because of low accuracy rate. On the other hand, Naïve Bayes was selected, although the processing time was more as compared to others but it yielded high accuracy rate which could not be ignored.

Step 3: Ensemble model tested with voting technique to classify the emails and check the accuracy. Training to testing ratio used in 20:80. Result shown in table 2.

4.2 Ensemble Model Testing

The Ensemble model has been created and tested using Rapid Miner software, version 5.3.015. Following results have been achieved;

Table 2: Ensemble Model Results

Parameter	Result
Accuracy	96.93%
Classification Error	3.07%
Precision	96.16%
Recall	98.8%
F measure	97.5%
False Positive (FP)	88.00
False Negative (FN)	25.00
True Positive (TP)	2205
True Negative (TN)	1362
False Positive rate (FPR)	0.06
True Positive Rate (TPR)	0.98
Area Under Curve (AUC)	0.983

Parameters detail is given in table 3.

Table 3: Parameter detail

Parameter	Formula
Accuracy	(TP+TN)/(P+N)
Classification Error	100 – Accuracy (%)
Precision	TP/(TP + FP)
Recall	TP/ (TP + FN)
False Positive (FP)	Absolute number of negative instances that were incorrectly classified as positive instances
False Negative (FN)	Absolute number of positive instances that were incorrectly classified as negative instances
True Positive (TP)	Absolute number of positive instances that were correctly classified as positive instances
True Negative (TN)	Absolute number of negative instances that were correctly classified as negative instances
False Positive rate (FPR)	FP/(FP+TN)
True Positive Rate (TPR)	TP/ (TP + FN)
Area Under Curve (AUC)	FPR(x-axis) vs TPR(y-axis)

ISBN: 978-1-941968-02-4 ©2014 SDIWC

Confusion Matrix

Table 4: Confusion Matrix

		Predicted	
Actual	True	**1.0**	**0.0**
	1.0	1362	25
	0.0	88	2205

Area Under Curve (AUC)

AUC or ROC curve is the graphical plotting of the performance of the classifier as its thresholds varied. The graph is plotted by placing the value of False Positive Rate (FPR) on the x-axis and True Positive Rate (TPR) on the y-axis.

FPR and TPR are calculated using following equations.

$$TPR = TP/ (TP + FN) \quad (1)$$

$$FPR = FP/ (FP + TN) \quad (2)$$

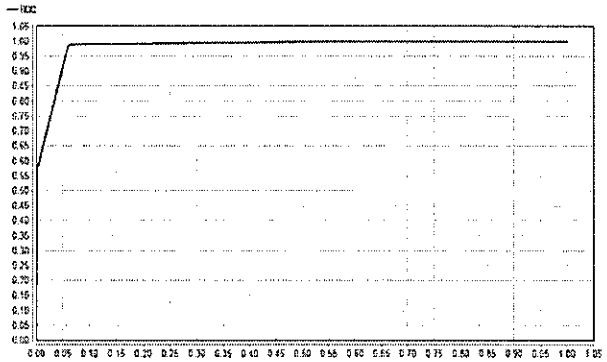


Figure 3: Area Under Curve

5 COMPARISON TO EXISTING CLASSIFICATION TECHNIQUES

In Table 5, comparison has been shown of our ensemble with the work of other researchers using other techniques for email classification. It can be seen that our ensemble has the highest accuracy as compared to other techniques.

Table 5: Comparison Results

Approach	Year	Accuracy	Note
An email classification Model based on Rough Set Theory [57]	2005	92.07%	20:80 Training to Testing ratio
Efficient Spam Email Filtering using Adaptive Ontology [54]	2007	96%	Applied on data set of 200 emails
E-mail Spam Classification With Artificial Neural Network and Negative Selection Algorithm [53]	2010	94%	Training to Testing Ratio is 60:40
Spam Classification Using Machine Learning Techniques [55]	2010	95%	Dataset of 2200 emails
Comparative Study on Email Spam Classifier Using Feature Selection Techniques [56]	2014	94.29%	Based on 28 features
Our Approach	2014	96.93%	20:80 Training to Testing ratio

When compared with the existing techniques, our approach used two new steps for processing and classifying the emails. Firstly we randomly checked the accuracy of different classifiers with 10:90 ratio training to testing dataset. During this process, the classifiers that took more time for processing were not considered for further testing like rule-based, neural net training algorithms. As processing times play a vital part in classification, therefore only those algorithms were selected that yielded low processing times.

Secondly, our approach uses the Majority voting based classification. This operator uses majority vote on top of the predictions of classifiers provided in its shell. This makes our approach comparable with Saeedian and Beigy [19] approach. In this approach an ensemble is proposed based on the combination of

clustering and weighted voting, The incoming email is compared with all defined clusters and hence similarity is checked to obtain weight for the final decision of the model. However there is extra pre-processing involved in this approach, for every new sample, the model compares it with all cluster centroids identifies its similarity to the cluster.

6 CONCLUSIONS AND FUTURE WORK

6.1 Conclusion

In this research work, we carried out a detailed literature review about the spam email classification techniques and proposed a different approach for acquiring relatively higher accuracy for classification. The variation of the proposed appraoach with other existing techniques lies in choosing different classification algorithms and Majority Voting method for the research problem (spam email classification).

The proposed ensemble model is validated by rigorously testing it in the software tool Rapidminer. Email dataset of about 4600 emails and 58 attirbutes was selected for this purpose. By keeping the ratio of 20:80 for training to testing data, the model yielded accuracy of about 97%. Detailed results when compared with other existing approached showed an improvement in the email classification.

6.2 Future Work

The following future work can be done on the basis of this project:

- Combine the proposed ensemble model with the pre-processing steps for the textual mining of the emails.
- To achieve higher accuracy, introduce more header filters and content based filters.
- Fuzzy logic can be used to distinguish spam at higher rate but a lot of complexity is involved in this.

7 REFERENCES

[1] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: An Efficient Data Clustering Method For Very Large Databases," Technical Report, Computer Sciences Department, University of Wisconsin-Madison, 1996.

[2] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya, "Phishing email detection based on structural properties," in Proc. 9th Annual NYS Cyber Security Conference., Jun. 2006.

[3] Noie, "SPAM REVIEW Interim Report by NOIE", Internet Society of Australia, A Chapter of the Internet Society, ABN 36 076 406 80, 2002.

[4] J. Goodman, G. V. Cormack, and D. Heckerman, "Spam and the ongoing battle for the inbox", Communications of the ACM, 50 (2):25{33}, 2007.

[5] M. Siponen and C. Stucke, "Effective anti-spam strategies in companies", An international study, In Proceedings of HICSS 2006, vol. 6, 2006.

[6] X. Wang and I. Cloete, "Learning to classify email: a survey", In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, ICMLC 2005, pp. 5716-5719, 2005.

[7] D. D. Lewis, and M. Ringuette, "Comparison of two learning algorithms for text categorization", In Proceedings of SDAIR, pp. 81–93, 1994.

[8] D. Koller, and M. Sahami, "Hierarchically classifying documents using very few words", In Machine Learning: Proceedings of the Fourteenth International Conference, pp. 170–178, 1997.

[9] J. D. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the poor assumptions of naïve bayes text classifiers". In Proceedings of ICML, pp. 616–623, 2003.

[10] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam", 2000.

[11] T. Fawcett, "In vivo, spam filtering: A challenge problem for kdd", In SIGKDD Explorations, vol. 5(2), pp. 140–148, 2003.

[12] X. Carreras, L. Marquez and J. G Salgado ,"Boosting trees for anti spam filtering", In International conference on Recent Advances in Natural Language Processing, pp. 58-64, 2001.

[13] E. Allman, "Spam Solution: Make the Spammers Pay", Article 2, 2003.

[14] T.A. ALMEIDA, J. ALMEIDA and A. YAMAKAMI, "Spam Filtering: How the Dimensionality Reduction Affects the Accuracy of Naive-Bayes Classifiers", Journal of Internet Services and Applications, 1(3), 183-200, London , February 2011.

[15] A. Khorsi, "An overview of content-based spam filtering techniques", Informatica; vol. 31, Issue 3, pg. 269, October 2007.

[16] T. K. Ho, "Random Decision Forest", Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, pp 278–282, 14–16 August 1995.

[17] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[18] F. S. Mehrnoush and B. Hamid, "Spam detection using dynamic weighted voting based on clustering," in Proceedings of the 2nd International Symposium on Intelligent Information Technology Application, (IITA '08), pp. 122–126, Shanghai, China, December 2008.

ISBN: 978-1-941968-02-4 ©2014 SDIWC