# acbpp actis 2014

Osaka

Official Conference Proceedings

ISSN: 2189 - 1028

"To Open Minds, To Educate Intelligence, To Inform Decisions"

The International Academic Forum provides new perspectives to the thought-leaders and decision-makers of today and tomorrow by offering constructive environments for dialogue and interchange at the intersections of nation, culture, and discipline. Headquartered in Nagoya, Japan, and registered as a Non-Profit Organization (一般社団法人) , IAFOR is an independent think tank committed to the deeper understanding of contemporary geo-political transformation, particularly in the Asia Pacific Region.

INTERNATIONAL

INTERCULTURAL

INTERDISCIPLINARY

iafor

# The Executive Council of the International Advisory Board

IAB Chair: Professor Stuart D.B. Picken  IAB Vice-Chair: Professor Jerry Platt

**Mr Mitsumasa Aoyama**
Director, The Yufuku Gallery, Tokyo, Japan

**Professor David N Aspin**
Professor Emeritus and Former Dean of the Faculty of Education, Monash University, Australia
Visiting Fellow, St Edmund's College, Cambridge University, UK

**Professor Don Brash**
Former Governor of the Reserve Bank, New Zealand
Former Leader of the New National Party, New Zealand
Adjunct Professor, AUT, New Zealand & La Trobe University, Australia

**Lord Charles Bruce**
Patron, Japan Society of Scotland
Patron, Scottish Centre of Tagore Studies at Edinburgh Napier University
Chairman, Kolkata Scottish Heritage Trust

**Professor Judith Chapman**
Professor of Education, Australian Catholic University, Australia
Visiting Fellow, St Edmund's College, Cambridge University, UK
Member of the Order of Australia

**Professor Chung-Ying Cheng**
Professor of Philosophy, University of Hawai'i at Manoa, USA
Editor-in-Chief, The Journal of Chinese Philosophy

**Professor Steve Cornwell**
Professor of English and Interdisciplinary Studies, Osaka Jogakuin University, Osaka, Japan
Osaka Local Conference Chair

**Professor Michael A. Cusumano**
SMR Distinguished Professor of Management and Engineering Systems, MIT Sloan School of Management, Massachusetts Institute of Technology, USA

**Professor Dexter Da Silva**
Professor of Educational Psychology, Keisen University, Tokyo, Japan

**Professor Georges Depeyrot**
Professor and Director of Research & Member of the Board of Trustees
French National Center for Scientific Research (CNRS) & L'Ecole Normale Superieure, Paris, France

**Professor Donald E. Hall**
Herbert J. and Ann L. Siegel Dean
LeHign University, USA

**Professor June Henton**
Dean, College of Human Sciences, Auburn University, USA

**Professor Michael Hudson**
President of The Institute for the Study of Long-Term Economic Trends (ISLET)
Distinguished Research Professor of Economics, The University of Missouri, Kansas City

**Professor Koichi Iwabuchi**
Professor of Media and Cultural Studies & Director of the Monash Asia Institute, Monash University, Australia

**Professor Sue Jackson**
Professor of Lifelong Learning and Gender & Pro-Vice Master of Teaching and Learning, Birkbeck, University of London, UK

**Professor Sing Kong Lee**
Director, The National Institute of Education, Singapore

**Professor Sir Geoffrey Lloyd**
Senior Scholar in Residence, The Needham Research Institute, Cambridge, UK
Fellow and Former Master, Darwin College, University of Cambridge
Fellow of the British Academy

**Professor Keith Miller**
Orthwein Endowed Professor for Lifelong Learning in the Science, University of Missouri-St.Louis, USA

**Professor Kuniko Miyanaga**
Director, Human Potential Institute, Japan
Fellow, Reischauer Institute, Harvard University, USA

**Professor Dennis McInerney**
Chair Professor of Educational Psychology and Co-Director of the Assessment Research Centre
The Hong Kong Institute of Education, Hong Kong SAR

**Professor Ka Ho Joshua Mok**
Chair Professor of Comparative Policy, Associate Vice-President (External Relations)
Dean of the Faculty of Arts and Sciences, The Hong Kong Institute of Education, Hong Kong SAR

**Professor Michiko Nakano**
Professor of English & Director of the Distance Learning Center, Waseda University, Tokyo, Japan

**Professor Baden Offord**
Professor of Cultural Studies and Human Rights & Co-Director of the Centre for Peace and Social Justice
Southern Cross University, Australia

**Professor Frank S. Ravitch**
Professor of Law & Walter H. Stowers Chair in Law and Religion, Michigan State University College of Law

**Professor Richard Roth**
Senior Associate Dean, Medill School of Journalism, Northwestern University, Qatar

**Professor Monty P. Satiadarma**
Clinical Psychologist and Lecturer in Psychology & Former Dean of the Department of Psychology and Rector of the University, Tarumanugara University, Indonesia

**Mr Mohamed Salaheen**
Director, The United Nations World Food Programme, Japan & Korea

**Mr Lowell Sheppard**
Asia Pacific Director, HOPE International Development Agency, Canada/Japan

**His Excellency Dr Drago Stambuk**
Croatian Ambassador to Brazil, Brazil

**Professor Mary Stuart**
Vice-Chancellor, The University of Lincoln, UK

**Professor Gary Swanson**
Distinguished Journalist-in-Residence & Mildred S. Hansen Endowed Chair, The University of Northern Colorado, USA

**Professor Jiro Takai**
Secretary General of the Asian Association for Social Psychology & Professor of Social Psychology Graduate School of Education and Human Development, Nagoya University, Japan

**Professor Svetlana Ter Minasova**
President of the Faculty of Foreign Languages and Area Studies, Lomonosov Moscow State University

**Professor Yozo Yokota**
Director of the Center for Human Rights Affairs, Japan
Former UN Special Rapporteur on Myanmar

**Professor Kensaku Yoshida**
Professor of English & Director of the Center for the Teaching of Foreign Languages in General Education, Sophia University, Tokyo, Japan

The International Advisory Board is composed of distinguished academics, business executives, former and current government officials, and community leaders. Its role is to consult and provide counsel on the business and affairs of IAFOR, and suggest areas of scholarly investigation. Membership of the International Advisory Board is by invitation only. For a full list please see the IAFOR website.

The Asian Conference on Technology, Information & Society 2014

Osaka, Japan

Official Conference Proceedings

ISSN: 2189-1028


The International Academic Forum (IAFOR)
Sakae 1-16-26-201
Naka Ward, Nagoya, Aichi
Japan 460-0008
ww.iafor.org

# Table of Contents

# Analysis of Time Series Mining in Manufacturing Problems

Ruhaizan Ismail, University Kebangsaan Malaysia, Malaysia
Zalinda Othman, University Kebangsaan Malaysia, Malaysia
Azuraliza Abu Bakar, University Kebangsaan Malaysia, Malaysia

The Asian Conference on Technology, Information & Society 2014
Official Conference Proceedings

**Abstract**
Recently, almost all modern manufacturing processes use operation that rely on automatic tools. The automatic tools react by material used and physical events. Physical events are determined by data patterns that can be definable by domain experts. However, when problems have not been predefined, it could cause critical errors especially in manufacturing plants. Furthermore, with machine complexity and new technology, this leads to a huge amount of data. These problems call for mining extraction to useful knowledge or patterns. Time series mining is about solving these problems by applying various approaches that suitable with manufacturing data. Until recently, the detection of time series data has received much less attention because of time series databases are usually very large, high dimensionality and the concepts of similarity can be subjective. The similarity may depends on the user, the domain, and the task at hand. As a result, a suitable time series mining approaches can be used to find any interesting and useful patterns. This study intents to present a broad classification of various time series problems in manufacturing operations. Also, the theoretical developments and analysis of the current available approaches used in time series mining are reviewed.

Keywords: data mining, time series, manufacturing, review

iafor
The International Academic Forum
www.iafor.org

## Introduction

Time series data appears naturally in almost fields of natural and social science as well as in numerous other disciplines. Most people are familiar with financial time series data like daily stock prices. For economist, they may want to identify the trend of changes in an annual household income over time and the relationship between different time series such as annual household incomes from different regions. In meteorological research, time series data can be mined for predictive analysis such as monthly average temperature (Zyu, 2004). Time series data also contain a valuable hidden knowledge within design and manufacturing environment. Time series databases consist of sequences of values or events obtained over repeated measurements of time. The values are typically measured at equal time intervals (e.g., hourly, daily, weekly). Thus, database contains time series data is called a sequence database. In a sequence database, each data is associated with the sequence of ordered events, with or without concrete notions of time (Han & Kamber, 2006).

This paper reviews on time series mining in manufacturing problems that apply various suitable approaches to manufacturing data. Until recently, time series mining has received much less attention because of time series databases are usually very large, high dimensionality and the concept of similarity can be subjective. The similarity may depends on the user, the domain, and the task at hand. As a result, a suitable time series mining can be used to find an interesting and useful data patterns in manufacturing operation. This study intents to present a broad classification of various time series manufacturing problems.

## 1. Time series mining

Temporal databases capture time-related attributes whose value change with time, for example, stock exchange data. There are five types of temporal data, which are static, sequences, time stamped, time series and fully temporal. However, two types of temporal data are dominant in the development of temporal data mining, i.e time-series data and sequence data. Time series data is defined as a time-ordered sequence of observation values of a physical or financial variables made at equally spaced time intervals (Palit & Popovic, 2006). It also can be a stretch of values on the same scale indexed by a time-like parameter, such as a range over the positive and negative integers or all real number or subsets of these. Also, the data is defined as a sequence of a real number that varies with time like stock prices, exchange rates, biomedical measurements data and so forth (Hsu et al., 2008). Another definition from Fu (2011), time series data is a collection of observations made by chronologically.

For sequence data, it is defined as a list of transaction, and transaction time is associated with each transaction (Hsu et al., 2008), or a sequence is an ordered list of event. An event can be represented as a symbolic value, a numerical, a vector or a complex data type. According to Xing et al. (2010), sequence data has subtypes that consists of a simple symbolic sequence, a complex symbolic sequence, a simple time series and a multivariate time series. A simple symbolic sequence is an ordered list of the symbols from the alphabet. However, the sequence like DNA sequence or protein sequence is called a complex symbolic sequence. It is complex because it is an ordered list of vector. Meanwhile a simple time series is a sequence of real values in time stamp from $t_1$ to $t_n$. However, when the simple time series used, the data needs to be transformed into symbolic sequence through discretization or symbolic

transformation. Meanwhile, a multivariate time series is a sequence of numerical vectors. It has been used for gesture and motion recognition (Xing et al., 2010).

Brillinger (2000) stated there are two categories of problem arise in time series data. There are scientific and statistical problem. The scientific problem is a problem that related with a systematic approach. It is designed to eliminate bias in the search of validation of facts. An example of scientific problems are measurement of uncertainty, difficulty in constructing a relationship, and spatially aggregated data (Brillinger, 1994). The systematic approaches that deal with scientific problems are smoothing, prediction, and association**.** However, more researches focus on solving problems related to predictive and associative approaches. This because of its related to forecasting the future values. Whereas, the statistical problem is a problem that associated with facts gathering and data analysis related to the process as a whole. There are example problems such as hidden frequencies, uncertainty computation, goodness of fit and testing. Beside these two problems, there are also special difficulties arise, such as missing values, censoring, measurement error, irregular sampling, feedback, outliers, shocks, signal-generated noise, trading days, festivals, changing seasonal pattern, measurement error, aliasing, and data observed in two series at different time points.

From the database perspective, time series data are facing problems such as large in data size, high dimensionality dan necessary to update continuously (Fu, 2011). In manufacturing, the large amount of data happens when the data is generated and collected during hourly or daily operations. Data through time, process and event are collected and stored in the database at various stages of design and production (Braha, 2001). This data may relate with many parameters such as products, materials, design and processes. The data contain hundreds of attributes, and the domain for each attribute can be large. This enormity is called high dimensionality data (Agrawal et al., 1998). From Esling & Agon (2012), time series data is essentially a high dimensional data. The high dimensionality problem makes the data difficult to be represented, imbalance proportion, long time scale and hinder in useful knowledge extraction. Because of the continuously updated characteristic, time series data also being called as dynamic or transactional data and it difficults to be captured. This nature makes the time series data mining more challenging. Time series researches are categorized into few different areas as:

i. Representation and indexing
   This task aims for dimensionality reduction. There are several methods in this task that representing time and frequency domains. In time domain approaches, the simplest and earlier developed method is sampling (Astrom, 1969). From sampling, the enhanced methods are developed by applying the mean value, approximating straight line, preserving the essential points or converting numeric time series to symbolic form. Representing time series in a frequency domain also a popular transformation technique. The proposed techniques are such as discrete Fourier transforms (Agrawal et al., 1993), similarity-based queries (Rafiei & Mendelzon, 2000), likelihood ratio statistics (Janacek et al., 2005) and discrete wavelet transforms (Chan & Fu, 1999). Other representing time series methods are hidden Markov models (Azzouzi & Nabney, 1998) and multidimensional indexing structure.

ii. Similarity measure

Similarity measure is like an exact match that based on the simple database definition. In similarity search, it finds data sequences that slightly differ from the given query sequence. Using an index to retrieve the sequence the actual distance between sequences is computed and discard any false matches. With the concept of matching, the similarity measure is represented in two ways, i.e. a whole sequence matching and subsequence matching. In the whole sequence matching, the query sequence is compared to each candidate in the series with the smallest distance. Example of the approaches are Euclidean distance, a constraint-based similarity query (Goldin & Kanellakis, 1995), pattern recognition (Morrill, 1998), dynamic time warping distance measure (Berndt & Clifford, 1994), threshold-based distance function (Abfalg et al., 2006) and parameter-light distance measure (Keogh et al., 2007). For subsequence matching, it's comparing the query sequence with the subsequence in the longer time series. The query sequence is required to be placed at every offset within the longer time series. The proposed subsequence approaches are window ordering (Kim & Jeong, 2007), linear segments (Morinaka et al., 2001), dynamic time warping with a suffix tree (Gusfield, 1997), and minimum-distance matching-window pair (Han et al., 2007). Also, a graph, framework or language can represent the relationship between subsequences and pattern query.

iii. Segmentation

Segmentation is a practice of dividing knowledge or information into groups that are similar in specific ways. It is considered as a discretization problem in a processing step. Earlier, a fixed length window has been proposed in order to segment data into a subsequence. There are two major disadvantages in using this approach, where first, the time series appeared meaningful pattern in different lengths, and second, meaningful pattern may be lost during cutting points. In order to prevent these problems, segmentation is handled by using a dynamic approach which identifies time points in more flexible way, i.e. the concept of different window widths. From that, more techniques have been proposed such as fuzzy clustering based segmentation (Abonyi et al., 2003), piecewise generalized likelihood ratio (Wang & Willett, 2004), and sliding test window (Chu, 1995). On the other hand, another way of solving the segmentation problem is by finding cyclic periodicity for all of the segments.

iv. Visualization

Visualization is one of the essential tools/ways in order to present the processed time series for further analysis. Many tools have been developed by past researchers such as by using pattern-based analysis (Schreck et al., 2007), and querying (Keogh et al., 2002). Also, other visualization techniques have been proposed such as a tree i.e. by converting each subsequence into a symbol string (Lin et al., 2005), bitmap (Kumar et al., 2005) and dot plots (Yankov et al., 2005).

v. Mining

Essentially, mining is aimed to discover hidden information or knowledge from either the original or transformed data. There are four issues arise in mining; pattern discovery, classification, rule discovery and summarization.

Pattern discovery is one of data mining task that mining various kinds of pattern, sequential patterns or sub-graph patterns. Another task in data mining is classification. Classification in time series mining usually transformed time series data into sequence data. From that, the task used to generate classifiers to find patterns and mining time series data into a useful result. For rule discovery, it is a task that mainly focus on symbolic items present in transactions. Rules are generated to describe each segment and in the terms of human readable. The advantages of rules, it can be explained clearly and tweak the underlying behavior. In summarization task, it is more on producing a compact description based on short or long-term time series data collection.

## 2. Issues in manufacturing

In most sectors, manufacturing is extremely competitive, and the financial margins that differentiate between success and failure are very tight, with most established industries needing to compete, produce and sell at a global level. Master to these trans-continental challenges, a company must achieve low-cost production. Still, it can maintain highly skilled, flexible and efficient workforces who can consistently design and produce high quality and low products (Choudhary et al., 2008).

There are various researches that review and focus on data mining in different application areas of manufacturing. The areas are design engineering, manufacturing systems, decision support system, shop floor control and layout, fault detection and quality improvement, maintenance, and customer relationship management. Choudhary et al. (2008), reviewed the various past researchers for data mining in manufacturing using five data mining functions; concept description, classification, clustering, prediction and association. They also categorized these functions for the manufacturing domains into quality control, job shop scheduling, fault diagnosis, manufacturing process, manufacturing system, maintenance, defect analysis, etc. Meanwhile, in Shanawaz et al. (2011), they presented an overview of eight techniques of temporal data mining; association, prediction, classification, clustering, characterization, search and retrieval, pattern discovery and trend analysis. As for Fu (2011), a review about time series in data mining, which was characterized into five tasks that already elaborate in section 2; representation and indexing, similarity measure, segmentation, visualization and mining. With focusing only in mining tasks, this paper highlights the past researchers in the manufacturing domain with pattern discovery issues.

From Fu (2011), pattern discovery is one of the issues that involves finding existing and surprising patterns. Pattern discovery also known as motif discovery, anomaly detection or finding discords. Pattern mining usually involved with temporal data because it implies the existence of time. There are three types of analysis that used in temporal data, such as temporal association rules, sequential pattern, and periodic patterns (Hsu et al., 2008). Temporal association rules regularly sampled univariate time series data. However, sequential pattern usually sampled in multivariate time series data. Sequential pattern concerned with finding precedence relationships. This pattern ordered association among data examples in the sequence. Some commonly used algorithms are Apriori, GSP, PrefixSpan and SPADE. Whereas, periodic pattern is considered as temporal regularity. This is usually used in web usage recommendation, weather prediction, computer networks and biological data (Huang & Chang, 2004).

With a focus on pattern discovery approaches, past researchers have reviewed several common problems in manufacturing operations.

***Database/product fault***

Zaki et al. (2000) presented an algorithm to extract patterns of events that predict failures in databases of plan executions. Analyzing execution traces is appropriate for planning that contain uncertainty, such as incomplete knowledge of the world or action with probabilistic effects. The causes of plan failures were extracted to feed the discovered patterns back into the planner. The goal is to find "interesting" sequences that have a high confidence of predicting plan failure. SPADE was used to mine such patterns.

Fountain et al. (2003), described a decision analysis for problem in integrated circuits manufacturing. The purpose of the testing was to avoid the expenses of packaging bad die and provide feedback by detecting die failures. They used a decision-theoretic approach to create a probabilistic pattern model of die failures. This model has been combined with a computation in deciding which die to test next and when to stop testing in real time.

Buddhakulsomsiri & Zakarian (2009) presented a sequential pattern mining algorithm that allow product and quality engineers to extract hidden knowledge from a large automotive warranty database. The algorithm used the elementary set concept and database manipulation techniques to search for patterns or relationships among occurrences of warranty claims over time. The sequential patterns represented in the form of association rules. The generated rules include the quality or warranty problems, and labor codes that occurred at a later time. Once a set of unique sequential patterns was generated, the algorithm applied a set of thresholds to evaluate the significance of the rules. The significant of rules provided the knowledge about how many product failures that lead to the future product faults

***Maintenance***

Sodiya et al. (2005) implemented Data mining-based Intelligent Maintenance System (DIMS) algorithm using Visual Basic 6.0 (VB). The algorithm applied a modified Apriori algorithm for frequent pattern mining on database. It used to extract predictive information needed for maintenance purposes and to identify association rules among faults. This system was designed to provide a way for timely and sophisticated analysis of fault database for effective equipment maintenance.

***Quality control***

Research by Purintrapiban and Kachitvichy (2003) proposed a fractal dimension based classifier in quality control. The proposed classifier was used to detect any unnatural patterns in the process. Fractal dimension is a recent classifier for pattern classification. The classifier is an index for measuring the complexity of an object. The results showed the classifier was effective in detecting non-periodic patterns such as natural patterns, linear patterns, systematic variable, stratification, mixture and sudden shifts. However, this approach was not sensitive for detecting linear patterns with a small slope.

***Knowledge acquisition***

Irani et al. (1993) developed an expert system for predicting the result of future experiments under various conditions, i.e. noisy data and limited availability training data. The system used generalized ID3 algorithm to diagnose and optimize a reactive ion etching process. The expert system was built for a knowledge acquisition. The discovered pattern was consistent with the data and has the same expectation from expert.

In 2003, Woon et al. proposed a method called PDMiner to mine web logs efficiently. This method was developed by using tree structure, association rule mining and sequential pattern mining techniques. PDMiner discovers the relationship among parts, assemblies, documents and how people interact with one another.

## 3. Our current research

From Choudhary et al. (2008), association rule and clustering have been proven as an effective method for extracting common patterns. These two approaches have mostly been used in the manufacturing field. However, there are a lot of problems in manufacturing industry that cannot be solved with only these two methods. The challenges are harder when it involves classifying sequence data or extracting pattern or knowledge. Time series methods have the potential to process and mine complex manufacturing data.

In our research, one framework are proposed for knowledge extraction from production data. A predictive model will be build from the extracted patterns by using sequential pattern approaches. The model can be used for identifying and characterizing sequence families, mining order from sequences and distinguishing any interesting sequences. Algorithms from sequential pattern mining are suitable for a large dataset and greatly reduced the amount of subsequence candidates compared to other approaches. The algorithms extract any hidden knowledge using pseudo-projection from frequent items within the data. The hidden knowledge of total execution time, lead time and makespan can be used to imply machine utilization and workload. This knowledge can be used to minimize capacity loss during the planning or rescheduling process.

## 4. Conclusions

From this study, a suitable time series mining approach can be used to find any interesting and useful data patterns in manufacturing operations. We have presented an overview of techniques for time series in time series mining. With many interesting techniques of solving time series data, it has been shown to be useful in many applications, especially in manufacturing domain. Past researchers have been tried to build a knowledge model for time series data. If a model is successful in interpreting the observed time series, the future value can be predicted to hold in the future.

## Acknowledgements

The first author acknowledges the financial support provided under grant FRGS/1/2014/ICT02/UKM/02/1 Fakulti Teknologi Dan Sains Maklumat, Universiti Kebangsaan Malaysia.

## References

Aßfalg, J., Kriegel, H.-P., Kröger, P., Kunath, P., Pryakhin, A., & Renz, M. (2006). Similarity search on time series based on threshold queries. *Advances in Database Technology-EDBT 2006*, 276-294: Springer.

Abonyi, J., Feil, B., Nemeth, S. & Arva, P. (2003). Principal Component Analysis Based Time Series Segmentation-Application to Hierarchical Clustering for Multivariate Process Data. *IEEE international conference on computational cybernetics,* 29-31.

Agrawal, R., Faloutsos, C. & Swami, A. (1993). Efficient Similarity Search in Sequence Databases. Springer.

Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P. (1998). Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. ACM.

Åström, K. J. (1969). On the Choice of Sampling Rates in Parametric Identification of Time Series. *Information Sciences*, *1(3),* 273-278.

Azzouzi, M. & Nabney, I. T. (1998). Analysing Time Series Structure with Hidden Markov Models. *Proceedings of the IEEE Conference on Neural Networks and Signal Processing*, 402-408.

Berndt, D. J. & Clifford, J. (1994). Using Dynamic Time Warping to Find Patterns in Time Series. *KDD workshop*, 359-370.

Braha, D. (2001). Data Mining for Design and Manufacturing - Methods and Applications: Kluwer Academic Publisher.

Brillinger, D. R. (1994). Examples of Scientific Problems and Data Analyses in Demography, Neurophysiology, and Seismology. *Journal of Computational and Graphical Statistics, 3(1)*, 1-22.

Brillinger, D. R. (2000). Time Series: General. Int. Encyc. Social and Behavioral Sciences.

Buddhakulsomsiri, J. & Zakarian, A. (2009). Sequential Pattern Mining Algorithm for Automotive Warranty Data. *Computer and Industrial Engineering, 57(1).*

Chan, K.-P. & Fu, A.-C. (1999). Efficient Time Series Matching by Wavelets. Data Engineering, 1999. *Proceedings., 15th International Conference on*, 126-133.

Choudhary, A. K., Harding, J. A. & Tiwari, M. K. (2008). Data Mining in Manufacturing: A Review Based on the Kind of Knowledge. *Journal Intelligent Manufacturing, 20(5)*, 501-521.

Chu, C.-S. J. (1995). Time Series Segmentation: A Sliding Window Approach. *Information Sciences, 85(1),* 147-173.

Esling, P. & Agon, C. (2012). Time-Series Data Mining. *ACM Computing Surveys (CSUR), 45(1),* 12.

Fountain, T., Dietterich, T. & Sudyka, B. (2003). Data Mining for Manufacturing Control: An Application in Optimizing Ic Tests. *Exploring artificial intelligence in the new millennium,* 381-400.

Fu, T.-C. (2011). A Review on Time Series Data Mining. *Engineering Applications of Artificial Intelligence, 24(1)*, 164-181.

Goldin, D. Q. & Kanellakis, P. C. (1995). On Similarity Queries for Time-Series Data: Constraint Specification and Implementation. *Principles and Practice of Constraint Programming—CP'95*, 137-153.

Gusfield, D. (1997). Algorithms on Strings, Trees and Sequences. *Computer Science and Computational Biology*, Cambridge University Press.

Han, J. & Kamber, M. (2006). Data Mining Concepts and Techniques. Academic press. Morgan Kaufmann.

Han, W.-S., Lee, J., Moon, Y.-S. & Jiang, H. (2007). Ranked Subsequence Matching in Time-Series Databases. *Proceedings of the 33rd international conference on Very large data bases*, 423-434.

Hsu, W., Lee, M. H. & Wang, J. (2008). Temporal and Spatio-Temporal Data Mining. New York, IGI Publishing.

Huang, K.-Y., & Chang, C.-H. (2004). Mining periodic patterns in sequence data. *Data Warehousing and Knowledge Discovery*, 401-410: Springer.

Irani, K. B., Cheng, J., Fayyad, U. M. & Qian, Z. (1993). Applying Machine Learning to Semiconductor Manufacturing. *IEEE Expert, 8(1),* 41-47.

Janacek, G.J., Bagnall, A.J., & Powell, M. (2005). A likelihood ratio distance measure for the similarity between the fourier transform of time series. *Advances in Knowledge Discovery and Data Mining*, 737-743: Springer.

Keogh, E., Hochheiser, H., & Shneiderman, B. (2002). An augmented visual query mechanism for finding patterns in time series data. *Flexible Query Answering Systems*, 240-250: Springer.

Keogh, E., Lonardi, S., Ratanamahatana, C. A., Wei, L., Lee, S.-H. & Handley, J. (2007). Compression-Based Data Mining of Sequential Data. *Data Mining and knowledge discovery, 14(1),* 99-129.

Kim, S.-W. & Jeong, B.-S. (2007). Performance Bottleneck of Subsequence Matching in Time-Series Databases: Observation, Solution, and Performance Evaluation. *Information Sciences, 177(22),* 4841-4858.

Kumar, N., Lolla, V. N., Keogh, E. J., Lonardi, S. & Ratanamahatana, C. (2005). Time-Series Bitmaps: A Practical Visualization Tool for Working with Large Time Series Databases. *SDM*, 531-535.

Lin, J., Keogh, E. & Lonardi, S. (2005). Visualizing and Discovering Non-Trivial Patterns in Large Time Series Databases. *Information visualization, 4(2),* 61-82.

Morinaka, Y., Yoshikawa, M., Amagasa, T. & Uemura, S. (2001). The L-Index: An Indexing Structure for Efficient Subsequence Matching in Time Sequence Databases. *Proc. 5th PacificAisa Conf. on Knowledge Discovery and Data Mining*, 51-60.

Morrill, J. P. (1998). Distributed Recognition of Patterns in Time Series Data. *Communications of the ACM, 41(5),* 45-51.

Palit, A. K. & Popovic, D. (2006). Computational Intelligence in Time Series Forecasting. *Theory and Engineering Applications*, Springer.

Purintrapiban, U. & Kachitvichyanukul, V. (2003). Detecting Patterns in Process Data with Fractal Dimension. *Computers & industrial engineering, 45(4),* 653-667.

Rafiei, D. & Mendelzon, A. O. (2000). Querying Time Series Data Based on Similarity. *Knowledge and Data Engineering, IEEE Transactions, 12(5),* 675-693.

Schreck, T., Tekušová, T., Kohlhammer, J. & Fellner, D. (2007). Trajectory-Based Visual Analysis of Large Financial Time Series Data. *ACM SIGKDD Explorations Newsletter, 9(2),* 30-37.

Shanawaz, M., Ranjan, A. & Danish, M. (2011). Temporal Data Mining: An Overview. *International Journal of Engineering and Advanced Technology (IJEAT), 1(1),* 20-24.

Sodiya, A., Longe, H. & Ibrahim, S. (2005). Data Mining-Based Intelligent Equipment Maintenance System in Telecommunication Network. *Journal of Applied Computer Science, 13(2),* 29-38.

Wang, Z. J. & Willett, P. (2004). Joint Segmentation and Classification of Time Series Using Class-Specific Features. Systems, Man, and Cybernetics, Part B: Cybernetics. *IEEE Transactions, 34(2),* 1056-1067.

Woon, Y.-K., Ng, W.-K., Li, X. & Lu, W.-F. (2003). Efficient Web Log Mining for Product Development. Cyberworlds, 2003. *Proceedings. 2003 International Conference on*, 294-301.

Xing, Z., Pei, J. & Keogh, E. (2010). A Brief Survey on Sequence Classification. *ACM SIGKDD Explorations Newsletter, 12(1),* 40-48.

Yankov, D., Keogh, E. & Lonardi, S. (2005). Dot Plots for Time Series Analysis. Tools with Artificial Intelligence, ICTAI 05. *17th IEEE International Conference*,168.

Zaki, M. J., Lesh, N. & Ogihara, M. (2000). Planmine: Predicting Plan Failures Using Sequence Mining. *Artificial Intelligence Review, 14(6),* 421-446.

Zhu, Y. (2004). High Performance Data Mining in Time Series: Techniques and Case Studies. *Tesis Degree of doctor of Philosophy, Department of Computer science, New York University*.

iafor

# *Analysis of Uncertainty in Time Series Data: Issues and Challenges*

Nabilah Filzah Mohd. Radzuan, University Kebangsaan Malaysia, Malaysia
Zalinda Othman, University Kebangsaan Malaysia, Malaysia
Azuraliza Abu Bakar, University Kebangsaan Malaysia, Malaysia

The Asian Conference on Technology, Information and Society 2014
Official Conference Proceedings

**Abstract**
This paper reviews issues and challenges of uncertainty in time series data. The aim of uncertainty analysis is to determine the ways of how to deal with uncertain data in order to gain knowledge, fit low dimensional model, and do prediction. So as to build an efficient predictive tool, uncertainty in data could not be ruled out because it may bring important knowledge. Uncertainty information arises from different resources such as process uncertainty, model uncertainty or data uncertainty. In this paper, issues and challenges of these uncertainties in time series data will be discovered and how these issues could be solved by data mining techniques will be discussed. Frequent pattern mining algorithm through FP-growth, Apriori algorithm and H-mine are methods that could be used to investigate the existing of uncertainty data. Meanwhile, Euclidean distance, particle swarm optimization, Monte Carlo simulation, and regression are methods that could be compared as prediction methods. These methods have been implemented in many data types since early 1900s. Also, this paper shows results of the uncertainty detection test on time series data sets. The test aims to prove the existing of uncertainty in the data. This work will benefit in many application domains.

Keywords: uncertainty, uncertain time series, mining algorithm

## Introduction

Uncertainty is a basic feature of automatic and semi-automatic data processes (Keijzer, Keulen, & Dekhtyar, 2007). There are many solutions have been proposed to reduce uncertainty because of risks in losing relevant information and misleading results (Radzuan, Othman, & Bakar, 2013). Uncertainty exists in time series data. Time series data is known as a stretch of values on a similar scale, indexed by a time that occurs naturally in many application domains such as environmental, economic, finance, and medicine. The aim of time series analysis is to formulate time series data to gain knowledge, to fit low dimensional models, and to make predictions. In reality, time series also deals with uncertainty. An uncertain time series data is a non-negative and precisely different ways in a number of fields (Cloke & Pappenberger, 2009; Lykoudis, Argiriou, & Dotsika, 2010). Particularly, uncertain data refers to data in which the ambiguity on whether it really takes place or not exists, or data for that the attribute values are not ascertained with 100 percent probability (Hooshsadat & Za, 2012). The combination of uncertainties is significant (Cloke & Pappenberger, 2009; Jankovic, 2004; Lykoudis et al., 2010) and brings important knowledge.

However, there are challenges to deal with when it involves domains such as manufacturing and weather forecast. Among the challenges include limited observational basis for seasonal and long term prediction, accurate forecasting of weather that may poses danger to aviation, prediction of product yielded during production process when expected situations happen, and many more (Williams et al., 2008). Researchers in manufacturing have attempted to discover the appropriate techniques specifically for modelling and processing uncertain time series for temporal data (Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012a; Dallachiesa, 2011). The involvement of this modelling and processing for uncertain time series is significant because they deal with query efficiency for accurate results (Zuo, Liu, Yue, Wang, & Wu, 2011). Meanwhile, in weather forecasting, new discovered knowledge from uncertain time series could be used in weather prediction or precipitation for the benefits of human being.

Since 1900s, there have been many studies carried out in dealing with uncertain time series. However, they used different terms. In 1905, researchers determined the relationships between normal and abnormal embryo development of a frog (Morgan, 1905). The inconsistency in the segmentation process was regarded as uncertain time series in the development stages (Hume, 1911). Meanwhile, in 1985, researchers helped identifying wet spells of weather and assessed the unusualness of the recent episode of heavy precipitation in meteorology department even though the uncertain whether prolonged the dry spell, which stroke the lake levels down to much lower levels before the onset of the next severe wet spell (Karl & Young, 1985). These situations visualize that the uncertain time series is very helpful and useful when knowledge is extracted from the data sets. Consequently, it is a promising direction to explore for more knowledge extraction methods in uncertain time series mining.

Further, uncertain time series mining is believed to be able to avoid risks and help in making better daily decisions. Uncertain time series mining also can improve the quality of demand, and identify temporal patterns that emerge and persist. This paper briefly describes the analysis of uncertain time series through issues and challenges. In response to that, this paper showcases the existence of uncertainty in selected uncertain time series datasets. In the remaining parts of this paper, some related works

are discussed in Sec. 2. Then, the analysis results are exhibited in Sec. 3. Next, Sec. 4 discusses the techniques and benefits. Finally, the conclusion is drawn in Sec.5.

**Related Work**

Uncertainty has been explicitly indicated as one of the future challenges in many fields (Halevy & Ordille, 2006). The uncertainty presents in all data processes and methods whether realize or not. The characterizing feature of uncertainty and other early works using uncertainty theories is after probabilities is used to select matching and non-matching objects (Hayne & Ram, 1990). Therefore, the uncertainty generated during data processes is lost (Dey, Sarkar, & Society, 2002; Florescu, Koller, & Levy, 1997).

There are relationships between original series (certain) and uncertain (Haitao & Xiaofu, 2009; Russa & Andrews, 2010). Generally, a clean time series data (certain time series data) is chosen for experiment. A certain time series data is a proper time series data which has been corrected or the inaccurate records have been removed from dataset. The certain time series is extracted to represent the original uncertain time series (Zuo et al., 2011). Uncertain time series can be treated as positional uncertain vectors (Abfalg, Kriegel, Kr, & Renz, 2009).

Besides, uncertainty exists in a modelling process, in which it arises from a fundamental choice as seen in grid resolution and from the parameterization of unresolved processes at the grid scale (Angew et al., 2004). Also, a high uncertainty brings big impact on prediction of regional climate change (Hawkins & Sutton, 2009). In fact, a lack of model diversity can cause a limited range of projections in climate change (Pennell & Reichler, 2011). Meanwhile, the distinct sources of uncertainty in prediction include internal variability, model uncertainty or response uncertainty, and scenario uncertainty (Hargreaves, 2010; Hawkins & Sutton, 2009).

The uncertain time series has been explored extensively in recent years. Uncertainty can be due to data aggregation, privacy-preserving transforms, and error-prone mining algorithms (Dallachiesa et al., 2012a; Dallachiesa, 2011). As a result, they found that uncertainty information might appear on different reasons (Dallachiesa et al., 2012a). Predicting uncertain time series appears to be a serious problem, as the existing forecast of certain time series does not purely mirror the ability of predicting future decisions. Uncertain time series in prediction is believed can avoid risks and help in making better daily decisions.

As an example, uncertain data is created by several applications in data forecasting, as can be seen in weather precipitation predicting for meteorology department, or in manufacturing demand prediction, which both actually can gain benefits in handling future outcome. Uncertain time series is important in making predictions. It influences the changeable climate that provides more useful information. Then, important knowledge can be tackled from this changeable gap that exists in uncertain time series data, in which the uncertainty can provide better results in terms of quality and efficiency (Dallachiesa et al., 2012a; Dallachiesa, 2011; Zuo et al., 2011).

Hence, the determination of predicting uncertain time series should be noted as a serious action to improve the quality of yield. The limitation found from the analysis can be used as an opening of the experiment and aim for securing the limitation for

enhancing the prediction outcome. Previous studies have discovered some possible properties of uncertainty in dataset (see Table 1). Also, clarification of uncertainty in dataset is important in identifying the type of data, so that they are not simply neglected. In normal practice, the organizer will neglect any data that they perceive as 'error' without investigating uncertain data's properties.

Table 1: The Properties of Uncertainty in Dataset

| The properties |
|---|
| • non-negative<br>• loss value or null, truly different ways in a number of fields<br>• data aggregation<br>• privacy-preserving transforms<br>• error-prone mining<br>• positional uncertain vectors<br>• exist in the modelling process where it arises from fundamental choice<br>• and, from the parameterization of processes unsolved at grid scale.<br>(Abfalg et al., 2009; Angew et al., 2004; Cloke & Pappenberger, 2009; Dallachiesa, Nushi, Mirylenka, & Palpanas, 2012b; Dallachiesa, 2011; Lykoudis et al., 2010) |

Therefore, there is an initiative to implement a number of algorithms consecutively to detect the uncertainty in dataset. In regards to that, Uncertain Associative Classifier (UAC) method (Hooshsadat & Za, 2012) could be used. It is measured partly on its accuracy, in which the percentage of accuracy is calculated using rule-based classifier on datasets. It is modelled based on a direct mining of discriminative patterns for classifying uncertain data at the level of uncertainty. In conjunction to this, a previous study found that the accuracy of data can be determined by uHARMONY, DTU, and uRule methods through UCI datasets (Hooshsadat & Za, 2012). Then again, uncertainty information arises from different resources such as process uncertainty, model uncertainty or data uncertainty. Frequent pattern mining algorithms through FP-growth, Apriori algorithm and H-mine are methods that could be used to investigate the existing of uncertainty in data. Table 2 shown the approaches include the advantages and disadvantages.

Table 2: The Uncertain Data Approaches

| Apriori (UApriori) | FP-growth (FP-tree) | H-mine (UH-mine) |
|---|---|---|
| Apriori identify the frequent items in the database and extending them to larger item sets appear sufficiently often in the dataset.<br>- UApriori is an extended from Apriori Algorithm.<br>- Efficient by employing pruning method. | Efficient and scalable especially for dense dataset<br>- Loss of compression properties.<br>- Large number of false positive is generated.<br>- The elimination of dataset further affects the efficiency. | Efficient and scalable especially for uncertain dataset.<br>- Can avoid generating a large number of candidate itemsets.<br>- Reduce memory requirements.<br>- Best trade-off in terms of running time and memory usage. |

**Experiment and Result**

The experiment in this study focuses on identifying uncertainty in selected datasets. The uncertain data is used to prove especially the accuracy of each prediction so that these methods can be studied for time series data. The data has gone through a discretization process (a process of organizing the dataset in minimizing redundancy and dependency, and makes it more informative to use).The discretization process involves scale-selective discretization (SSD) procedure as in (Vuorinen, Larmi, Schlatter, Fuchs, & Boersma, 2012). This SSD separates small and large scales of the flow using a high-pass filter.

The Apriori is used as a generate-and-test approach by generating the dataset attributes and testing if they are frequent or not. Generation of dataset attributes are disconnected, where it is involve checking subset in each attributes and scanning multiple databases. Then, FP-Growth allows frequent attributes discovery without dataset attributes generation. There are two steps in this approach; first, it builds a compact data structure called FP-tree where it is built using two passes over the dataset. Second, it extracts frequent attributes directly from FP-tree where traversal through FP-tree. The H-mine tries to avoid generating a large number of dataset attributes and uses all involved attributes without eliminating or avoiding the null value.

The three approaches are intersect with Uncertain Associative Classifier (UAC) method as implemented in (Hooshsadat & Za, 2012). The UAC algorithm can only be implemented after the trained dataset goes through a discretization process. The UAC algorithm is visualized in Figure 1. It involves three stages of UAC rule filtering of the three approaches. Further, the algorithm of each stage is detailed in Appendix A. Briefly, the UAC algorithm selects one classifying rule for each instance which has the highest relative precedence with respect to the test instance (Hooshsadat & Za, 2012).

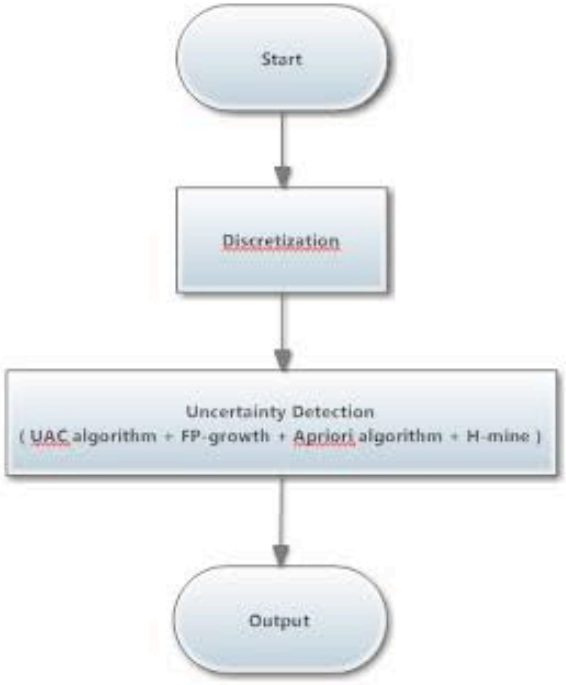



Figure 1: Flow of approaches

In this experiment, 43 uncertain time series datasets from UCI benchmark data were used. The results in Table 2 show the existence of uncertainty in the datasets. Particularly, to add the 10 percent uncertainty to an attribute, it is attached with a 0.9 probability and the remaining 0.1 is distributed randomly among other values appear in the domain (Hooshsadat & Za, 2012). Eventually, the highest percentage of uncertainty represents the highest uncertainness in the data.

Table 2 Result of uncertainty percentage detection for 43 dataset

| Dataset | Uncertainty (%) |
|---|---|
| Single Chest Mounted | 68.2 |
| ADLs | 24.9 |
| Amazon Access | 70.4 |
| ASL | 44.3 |
| ASL signs | 39.3 |
| Bach Chorales | 83.6 |
| Buzz | 28.1 |
| CalIt2 | 32.0 |
| Character Trajectories | 85.7 |
| DS Activities | 41.9 |
| Daphnet | 37.3 |
| Wrist worn | 21.5 |
| Diabetes | 12.4 |
| EEG | 20.0 |
| EEG Eye State | 25.5 |
| EMG Lower | 88.1 |
| GSA Drift | 0 |
| GSA turbulent gas mixtures | 23.8 |
| GSA under flow modulation | 48.0 |
| GSA open sampling settings | 26.0 |
| Gesture Phase Segmentation | 31.8 |
| Smartphones | 34.0 |
| ICU | 50.4 |
| Individual household | 10.8 |
| Istanbul SE | 39.8 |
| Japanese Vowels | 11.1 |
| Localization Data | 87.9 |
| Opportunity Activity | 42.3 |
| Ozone | 14.3 |
| PAMAP2 | 39.3 |
| PEMS-SF | 23.6 |
| Pioneer-1 | 28.1 |
| Predict keywords | 32.0 |
| Pseudo Periodic Synthetic | 85.7 |
| Realdisp Activity | 61.9 |
| Robot | 77.3 |
| SML2010 | 81.5 |
| Spoken Arabic Digit | 12.4 |
| Synthetic Control | 80.0 |
| URL Reputation | 23.0 |
| Walking Activity | 34.3 |
| Vicon | 75.0 |

The experiment explained in the previous paragraph proves the existence of uncertainty in the time series dataset. The time series dataset was normalized before implemented on UAC algorithm. Normalization is important in order to minimize redundancy and isolate data. The mining process of time series data differs from normal dataset as the data properties itself are different. The result of time series dataset is same with previous study (Hooshsadat & Za, 2012) where there is existence of uncertainty in the dataset.

**Discussion**

The yield of prediction and knowledge from uncertain data brings important meaning for future prediction especially in weather domain. In a real situation, unpredictable events happen without being anticipated. In this study, the reviewed methods bring benefit to domain in predicting the uncertain time series data. The performed analytical and experimental comparisons of techniques described in the previous section should be further experimented in order to get accurate prediction.

The FP-growth approach extracts frequent attributes from the FP-tree. The FP-tree can be built if only consider the transactions containing a particular attributes or else removing the attributes from all transactions. The H-mines approach help in minimizing items lost from that transactions. The compressed datasets have high tendency of losing attributes. All the uncertain properties in the datasets have been calculated and shown in percentage. From the percentage values, the uncertainty in datasets are detected.

In this study, there are differences between uncertain data and uncertain time series data. While uncertain data refers to static data (Aggarwal, Li, Wang, & Wang, 2009), uncertain time series data refers to continuous data (Gagne, McGovern, & Xue, 2011). However, both collected data often inaccurate and are based on incomplete or inaccurate information. The detection test on uncertainty has shown that there are uncertainties in the datasets that would bring highly potential in yielding information for future prediction. The test helps the organizer to not neglect any data that they perceive as ‘error’. Therefore, the UAC method could be utilized for time series data in determining uncertainty in data. Then, the yield, which is uncertain time series data, of the process can be implemented on prediction methods.

**Conclusion**

This paper explains on evaluation methods used in uncertain time series. The analysis on previous works and the experiments outline the methods on certain data in order to extract knowledge for future work. This study discovers that there are methods that bring limitation in their prediction processes. The presence of uncertainty in dataset can be determined through a combination of FP-growth, Apriori algorithm, H-mine and UAC method. The data first gone through a discretization process involving SSD, in which it is a process of organizing the dataset in minimizing redundancy and dependency, and make it more informative to use. Through the experiment, on uncertainty existence in uncertain datasets have proved that uncertainties exist in time series data. Although the experiment brings benefits to domain, still future actions must be taken in obtaining accurate prediction in uncertain time series.

**References**

Abfalg, J., Kriegel, H., Kr, P., & Renz, M. (2009). Probabilistic Similarity Search for Uncertain Time Series. In *SSDBM 2009 Proceedings of the 21st International Conference on Scientific and Statistical Database Management* (pp. 435 – 443). doi:10.1007/978-3-642-02279-1_31

Aggarwal, C. C., Li, Y., Wang, J., & Wang, J. (2009). FREQUENT PATTERN MINING ALGORITHMS WITH UNCERTAIN DATA. In *Managing and Mining Uncertain Data* (pp. 428–459).

Angew, M. R. S., Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., … Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature Publishing Group*, *430*(August 2004), 768–772. doi:10.1038/nature02770.1.

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3-4), 613–626. doi:10.1016/j.jhydrol.2009.06.005

Dallachiesa, M. (2011). Similarity Matching for Uncertain Time Series : Analytical and Experimental Comparison, 8–15.

Dallachiesa, M., Nushi, B., Mirylenka, K., & Palpanas, T. (2012a). Uncertain Time-Series Similarity : Return to the Basics. In *Proceedings of the VLDB Endowment* (Vol. 5, pp. 1662–1673).

Dallachiesa, M., Nushi, B., Mirylenka, K., & Palpanas, T. (2012b). Uncertain Time-Series Similarity : Return to the Basics, *5*(11), 1662–1673.

Dey, D., Sarkar, S., & Society, I. C. (2002). Generalized Normal Forms for Probabilistic Relational Data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *14*(3), 485–497.

Florescu, D., Koller, D., & Levy, A. Y. (1997). Using probabilitistic Information in Data Integration. In *Proceeding VLDB '97 Proceedings of the 23rd International Conference on Very Large Data Bases* (pp. 216–225).

Gagne, D. J., McGovern, A., & Xue, M. (2011). Machine learning enhancement of storm scale ensemble precipitation forecasts. In *Proceedings of the 2011 workshop on Knowledge discovery, modeling and simulation - KDMS '11* (p. 45). New York, New York, USA: ACM Press. doi:10.1145/2023568.2023581

Haitao, L., & Xiaofu, Z. (2009). Precipitation Time Series Predicting of the Chaotic Characters Using Support Vector Machines. In *2009 International Conference on Information Management, Innovation Management and Industrial Engineering* (pp. 407–410). Ieee. doi:10.1109/ICIII.2009.407

Halevy, A., & Ordille, J. (2006). Data Integration : The Teenage Years. In *VLDB '06 Proceedings of the 32nd international conference on Very large data bases* (Vol. 12, pp. 9–16).

Hargreaves, J. C. (2010). Skill and uncertainty in climate models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(4), 556–564. doi:10.1002/wcc.58

Hawkins, E., & Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1107. doi:10.1175/2009BAMS2607.1

Hayne, S., & Ram, S. (1990). Multi-user view integration system (MUVIS): an expert system for view integration. In *[1990] Proceedings. Sixth International Conference on Data Engineering* (pp. 402–409). IEEE Comput. Soc. doi:10.1109/ICDE.1990.113493

Hooshsadat, M., & Za, O. R. (2012). An Associative Classifier For Uncertain Datasets. In *Advances in Knowledge Discovery and Data Mining* (p. pp 342–353). doi:10.1007/978-3-642-30217-6_29

Hume, W. F. (1911). The Effects of Secular Oscillation in Egypt during the Cretaceous and Eocene Periods. *Quarterly Journal of the Geological Society*, *67*(1-4), 118–148. doi:10.1144/GSL.JGS.1911.067.01-04.05

Jankovic, V. (2004). Science Migrations: Mesoscale Weather Prediction from Belgrade to Washington, 1970–2000. *Social Studies of Science*, *34*(1), 45–75. doi:10.1177/0306312704040490

Karl, R. T., & Young, J. P. (1985). Recent heavy precipitation in the vicinity of the Great Salt Lake: just how unusual. *Journal of Climate and Applied Meteorology*, *25*, 353–363.

Keijzer, D. A., Keulen, V. M., & Dekhtyar, A. (2007). Report on the First VLDB Workshop on Management of Uncertain Data (MUD). *ACM SIGMOD Record*, *36*(4), 18–32. doi:10.1145/1361348.1361363

Lykoudis, P. S., Argiriou, A. A., & Dotsika, E. (2010). Spatially interpolated time series of δ18O in Eastern Mediterranean precipitation. *Global and Planetary Change*, *71*(3-4), 150–159. doi:10.1016/j.gloplacha.2009.09.004

Morgan, T. H. (1905). The relation between normal and abnormal development of the embryo of the frog. *Archiv Für Entwicklungsmechanik Der Organismen*, *19*(4), 581–587. doi:10.1007/BF02321200

Pennell, C., & Reichler, T. (2011). On the Effective Number of Climate Models. *Journal of Climate*, *24*(9), 2358–2367. doi:10.1175/2010JCLI3814.1

Radzuan, N. F. M., Othman, Z., & Bakar, A. A. (2013). Uncertain Time Series in Weather Prediction. *Procedia Technology*, *11*(Iceei), 557–564. doi:10.1016/j.protcy.2013.12.228

Russa, C. La, & Andrews, K. (2010). AC 2010-1093 : Managing A Digitization Project : Issues for State Agency Publications with Folded Maps. *American Society for Engineering Education*, 1–6.

Vuorinen, V., Larmi, M., Schlatter, P., Fuchs, L., & Boersma, B. J. (2012). A low-dissipative, scale-selective discretization scheme for the Navier–Stokes equations. *Computers & Fluids*, *70*, 195–205. doi:10.1016/j.compfluid.2012.09.022

Williams, K. J., Ahijevych D.A., Kessinger C.J., Saxen T.R., Steiner M., & Dettling S. (2008). A Machine Learning Approach to Finding Weather Regimes and Skillful Predictor Combinations for Short-Term Storm Forecasting. *National Center for Atmospheric Research*, 1–6. Retrieved from http://opensky.library.ucar.edu/collections/OSGC-000-000-003-270

Zuo, Y., Liu, G., Yue, X., Wang, W., & Wu, H. (2011). Similarity Matching over Uncertain Time Series. *2011 Seventh International Conference on Computational Intelligence and Security*, 1357–1361. doi:10.1109/CIS.2011.302

**Acknowledgement**

This work is acknowledged by Universiti Kebangsaan Malaysia (UKM), Malaysia through Research University (RU) grant scheme number of Project FRGS/1/2014/ICT02/UKM/02/1 for financial support.

## Appendix A

The UAC algorithm (Hooshsadat & Za, 2012).

Algorithm 1 UAC Rule Filtering: Stage 1

```
1:    Q = ;;U = ;;A = ;
2:    for all i 2 Dataset do
3:    i:ucRule = firstCorrect(i)
4:    i:cApplic = _(i:ucRule; i)
5:    i:uwRule = firstWrong(i)
6:    i:wApplic = _(i:uwRule; i)
7:    U:add(ucRule)
8:    ucRule:covered[i:class] + +
9:     if (ucRule _[i] uwRule) and ucRule _ uwRule
then
10:   Q:add(ucRule)
11:   flag(ucRule)
12:   else
13:   A:add(< i:id; i:class; ucRule; uwRule>)
14:   end if
15:   end for
```

Algorithm 2 UAC Rule Filtering: Stage 2

```
1:    RepDAG = ;
2:    for all < i:id; y; ucRule; uwRule>2 A do
3:    if flagged(uwRule) then
4:    ucRule:covered[y] □□
5:    uwRule:covered[y] + +
6:    else
7:    wSet = allCoverRules(U; i:id; ucRule)
8:    if !RepDAG:contains(ucRule) then
9:    RepDAG:add(ucRule)
10:   end if
11:   for all w 2 wSet do
12:   w:replace:add(<ucRule; i:id; y >)
13:   w:covered + +
14:   ucRule:incom + +
15:   if !w 2 RepDAG then
16:   RepDAG:add(w)
17:   end if
18:   end for
19:   Q = Q:add(wSet)
20:   end if
21:   end for
22:   S   set of all nodes with no incoming edges
23:   while S 6= ; do
24:   r = S:next() fnext removes a rule from the setg
25:   for all <ucRule; id; y >2 r:replace do
```

```
26:     if (r:covered[r:class] > 0) then
27:     if id is covered then
28:     r:covered[y] □□
29:     else
30:     ucRule:covered[y] □□
31:     Mark id as covered.
32:     end if
33:     end if
34:     ucRule:incom□□
35:     if ucRule:incom = 0 then
36:     S:add(ucRule)
37:     end if
38:     end for
39:     end while
```

Algorithm 3 UAC Rule Filtering: Stage 3

```
1:      C = ;
2:      for all r 2 Q do
3:      if r:covered[r:class] > 0 then
4:      finalSet:add(r)
5:      ruleErrors+ = computeError(r)
6:      defClass = addDefaultClass()
7:      defErrors = computeDefErr(defClass)
8:      defAcc = addDefAcc(uncovered(D) □defErrors)
9:      totalError = defErrors + ruleErrors
10:     C:add(r; totalError; defClass; defAcc)
11:     end if
12:     end for
13:     Break C from the rule with minimum error
14:     C contains the _nal set of rules
15:     default = defClass:get(C:size)
16:     defApplic = defAcc:get(C:size)
                        jTj
```

iafor

# *Cross-Cultural Study on the Value Structure of Mobile Internet Usage: Comparison between Indonesia and Thailand*

Chalita Srinuan, King Mongkut's Institute of Technology Ladkrabang, Thailand
Ranti Yulia Wardani, STIE YKPN, Yogyakarta, Indonesia

The Asian Conference on Technology, Information & Society 2014
Official Conference Proceedings 2014

**Abstract**
This study aims to examine a pattern of mobile Internet usage from perspective of intercultural among developing countries. Data obtained from a survey of questionnaire which is distributed in both Thailand and Indonesia during 2013. The multiple regression analysis has been applied to identify between value structures (functional value, emotional value, social value and monetary value) and overall satisfaction in each country. Finding shows that functional value, emotional value, social value, and monetary value significantly affect satisfaction for Indonesia users which is similar to Thai users except the emotional value does not affect the satisfaction. A comparison test to examine the difference in value structure between Indonesia and Thailand has been analyzed. The results show that the functional value, emotional value, social value and monetary value between Indonesia and Thailand are different. The results of these differences are influenced by the perceptions of respondents in Indonesia and Thailand to the different mobile Internet usage pattern. In this case it turns out different cultures influence the way in perceiving the use of different technologies, especially in the mobile Internet usage pattern.

Keywords: Cross- culture, Mobile Internet, Thailand, Indonesia

## Introduction

Over the past years, information technology and communications (ICT) continue to spread throughout the world. Internet is one of the important tools in IT that help people doing their business, job, and many activities. Internet has a strong ability to help people find information. Internet provides a wealth of information from thousands of online publications such as newspapers, journals, reports, entertainment, and data. In addition, there are many websites that allow us to connect with many people in our life by using social networking sites such as Facebook and Twitter. Internet is easy to access, and can assist in communicating with users around the world.

Internet would lead people to the need of access Internet via cellular networks every time and everywhere. People who travel from place to another place to do some activities prefer to get information quickly using mobile Internet. One of the tools that people need to be able to access the Internet quickly anywhere and anytime is the mobile Internet. Based on ITU (2009) access to the Internet via mobile has been growing rapidly align with the developing availability of network infrastructure and devices, including mobile handsets and data cards that allow users to access Internet via mobile Internet devices.

ITU (2012) stated that between 2010 and 2011, mobile-cellular subscriptions registered continuous double-digit growth in developing country markets. Number of mobile-cellular subscriptions increased by more than 600 million, almost all of them in the developing world, to a total of around 6 billion, or 86 per 100 inhabitants, globally. In developing countries, mobile sector competition is getting stronger. This condition encourage by strong growth of mobile Internet usage. Prepaid services in developing countries lead to communication accessible by many low-income society. Mobile services slowly change fixed lines services. Moreover, more than 70 per cent of all mobile cellular subscriptions are prepaid (and as many as 87 per cent in developing countries).

A sharp increase in mobile broadband subscriptions growth reflects a handheld device that can access the Internet via smartphones or tablet, coupled with the launch of high-speed networks and mobile services (3G and above) in developing countries. As the network operators are working to increase and expand the scope of services, they can achieve an increase in the proportion of urban and rural population of the world.

Mobile phones also penetrate in some social hierarchy such as lower class, middle class, and upper class. Mobile phone is not only using middle class and upper class market but also using by lower class market. Duncombe & Boateng (2009) stated that mobile phones are becoming increasingly part of the everyday lives of the poor, it is argued that they have potential to become a low-cost accessible account or delivery channel for financial information, services and transactions, thus facilitating innovations including micro-payments (m-payments), electronic money (e-money), and a mobile banking channel (m-banking). Thus, mobile phones function becoming important in every single class society. It will encourage people in every single class society to use mobile phone function more than phone calls and integrated with Internet.

Previous studies have been conducted to investigate the cross-cultural differences in different academic fields, such as psychology, philosophy, and human resources. In some studies are little attention given to cross-cultural research in the field of ICT because there are difficulties in explaining and measuring cultural concepts related to ICT (Straub et al, 2002). This study will use mobile Internet as one of the tool of ICT and measure it with the cultural concepts. People from different countries will have different adoption of the mobile Internet usage pattern. Therefore it would be interesting to investigate the mobile Internet usage pattern in two different countries. This study is based on the following research questions:

1. Do functional value, emotional value, social value, and monetary value affect mobile Internet service satisfaction in Indonesia and Thailand, respectively?
2. Is there any difference in value structures between Indonesia and Thailand?

## Literature review

Kalakota & Robinson (2002) are stated that mobile mostly used to imply that the device has an "always on" connection to the Internet. Moreover, Kalakota & Robinson (2002) explained that mobile phones with the Internet connections are commonly called wireless, thus implying that the experience is based on a real-time live Internet connection via satellite, cellular, or radio transmitters.

In this study mobile Internet can be measured with values structure construct of different culture. A consumption values consists of several components that affect the value of consumer choice behavior, for example: functional value, conditional value, social value, emotional value and value of knowledge (Sheth et al, 1991). In this study, the purpose of framework structure of value will be formulated in variety types of values that users try to satisfy by using the mobile Internet.

Kreitner and Kinicki (2007) stated that culture is the beliefs and values about how a community should do to perform actions. They also defined define a value is an enduring belief in the model of conduct or end-state. Structure values are analytical framework that consists of different types of value provided by the mobile Internet (Lee et al, 2002). Value can be recognized as a tool or a key parameter in the study to measure culture. Value not only can be used to understand human behavior, but also the value can be used to conduct business (Lee et al, 2002).

Value is a concept that consists of several sub-components that are heterogeneous (Sweeney & Soutar, 2001). There are four sub-values: functional value, emotional value, social value, and monetary value. The definition of functional value is functional or technical practical benefits that users can get from using the mobile Internet. Wardani & Warsono (2012) found that Japanese functional value is more likely to affect satisfaction and Indonesian social value is a high influential effect on satisfaction. Japanese respondents are exhibiting more utilitarian in pattern of mobile Internet usage. On the other hand Indonesian respondents are exhibiting more hedonic in pattern of mobile Internet usage. It can be explained from the result that shows Indonesian emotional value is significantly higher than Japanese emotional value. The emotional value is one of the indicators that people tend to use mobile Internet to fulfill their emotional need by accessing preferred website. Hedonic criteria are related to feelings and emotions such as fun, relaxation, or pleasure. This means that

Indonesian respondents perceived mobile Internet to fulfill their emotional need, which is higher than other values.

Emotional value is a mental or psychological need of mobile Internet users (Sweeney & Soutar, 2001). Arambewela, et al. (2005) examined differences in personal values among Asian postgraduate international students from China, India, Indonesia and Thailand who are studying in some universities in Australia. The study was conducted based on nationality and discuss the marketing implications of each difference. Personal value is one way which insights can be obtained from the students, especially those related to the needs and preferences. He found that Indonesia and Thailand college students have a tendency for greater importance to the values associated with hedonism.

Social value is social benefits gained by the user when they can connect to others via mobile Internet (Sheth et al, 1991). Indonesian society has high social network characteristics (Subagyo, 2009). Indonesian country is said to have high expression of family value. These models may include more than two hundred million people in Indonesia. Indonesian people who live together are in a spirit of kinship (Mulder, 2000). According to Nguyen (2005) Thai people really appreciate the friendship and tend to look for a permanent friendship. Thailand society almost has a similarity with other Asian cultures, Thailand worth more or less influenced by Confucianism. They are usually: devoted, respect for age, seniority and hierarchy, respect, dignity, ethical, true friendship, be averse to arrogance and vanity, have a preference to learn, and a belief in simplicity (Nguyen, 2005).

Monetary value means how satisfying mobile Internet service compared to the cost or time or effort which is spend in using of mobile Internet (Sweeney and Soutar, 2001). Monetary value should not be ignored because mobile Internet user must pay relatively high charges for mobile Internet services usage (Lee et al, 2002). According Information Telecommunication Union (ITU) in 2012, the ICT Price Basket (IPB) is a unique global measure that provides important information related to telecommunications and information technology services. The IPB consists of three components different prices: fixed line, mobile cellular and fixed broadband services. They are then calculated as a percentage of average gross national income (GNI) per capita. The IPB is the value calculated from the sum of the price of each sub-basket (in USD) as a percentage of a country's monthly GNI per capita divided by three. Calculated results by ITU indicated that cost of for mobile Internet services in Thailand and Indonesia are different both in Prepaid and Postpaid subscription. Thailand service providers offer a lower price plan of mobile internet prepaid subscription and higher price plan for postpaid subscription as compared to Indonesia service providers. Therefore, monetary value will affect their satisfaction using the mobile Internet unavoidably.

Satisfaction is the overall experience perceived by the customer from the beginning to the end use of the product or service usage of the product or service providers (Johnson & Fornell, 1991). Overall satisfaction using the mobile Internet is the overall value of mobile Internet service as experienced by the user. In this study satisfaction value is the dependent variable.

## Data and method

The survey method is used in this study as a data collection technique. Data has been collected from the survey which undertaken in Indonesia and Thailand. The questionnaire was given to respondents in Thailand and Indonesia directly. There are four parts in the questionnaire: the user's demographic questions, usage behavior, four value and satisfaction. Sample size of this study was 300 samples and was divided 150 samples for each country. The response rate was 91.33 percent. The independent variables in this study are value variables and dependent variable is satisfaction. Culture will be treated as moderating variable. Multiple regression analysis is used to identify the relationship between value structures and overall satisfaction in each country. Independent samples T-test is also used in this research to test how different value structures between the two countries.

## Results and discussion

The regression results show that the value structure (functional value, emotional value, social value, monetary value) has influence on the satisfaction of Indonesian mobile Internet user. Considering Thailand mobile Internet user, function value, social value and monetary value simultaneously affect the satisfaction of Thailand respondents. The emotional value is not significantly affecting the satisfaction of mobile Internet user as show in Table 1.

**Table 1.** Multiple Regressions

| Model | Unstandardized Coefficients | | Standardized Coefficients |
|---|---|---|---|
| | B | Std. Error | Beta |
| Indonesia | | | |
| (Constant) | 0.258 | 0.940 | |
| FVINA | 0.167 | 0.045 | 0.301**** |
| EVINA | 0.111 | 0.045 | 0.205** |
| SVINA | 0.082 | 0.032 | 0.153** |
| MVINA | 0.240 | 0.042 | 0.350*** |
| Thailand | | | |
| (Constant) | -0.022 | 0.943 | |
| FVTHAI | 0.215 | 0.046 | 0.371*** |
| EVTHAI | -0.022 | 0.045 | -0.038 |
| SVTHAI | 0.183 | 0.045 | 0.276*** |
| MVTHAI | 0.246 | 0.042 | 0.367*** |

*** , ** refers to significant level at 1% and 5% respectively

Independent samples T-test was also used in this study to test how different value structures between the two countries. Findings in Table 2 shows that there is functional value, emotional value, social value and monetary value difference between the respondent in Indonesia and Thailand

**Table 2.** Independent T-test

| | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F | Sig. | t | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| | | | | | | | | | Lower | Upper |
| FV | Equal variances assumed | 0.922 | 0.338 | 5.221 | 272 | 0.000 | 3.711 | 0.711 | 2.312 | 5.111 |
| | Equal variances not assumed | | | 5.204 | 265.042 | 0.000 | 3.711 | 0.713 | 2.307 | 5.115 |
| EV | Equal variances assumed | 0.897 | 0.345 | 7.167 | 272 | 0.000 | 5.146 | 0.718 | 3.733 | 6.560 |
| | Equal variances not assumed | | | 7.137 | 263.346 | 0.000 | 5.146 | 0.721 | 3.727 | 6.566 |
| SV | Equal variances assumed | 7.591 | 0.006 | 3.127 | 272 | 0.002 | 2.139 | 0.684 | 0.792 | 3.485 |
| | Equal variances not assumed | | | 3.094 | 245.603 | 0.002 | 2.139 | 0.691 | 0.777 | 3.500 |
| MV | Equal variances assumed | 0.223 | 0.637 | 5.153 | 272 | 0.000 | 3.067 | 0.595 | 1.895 | 4.238 |
| | Equal variances not assumed | | | 5.152 | 269.654 | 0.000 | 3.067 | 0.595 | 1.895 | 4.238 |
| SAT | Equal variances assumed | 0.576 | 0.448 | 6.165 | 272 | 0.000 | 2.482 | 0.403 | 1.690 | 3.275 |
| | Equal variances not assumed | | | 6.157 | 268.299 | 0.000 | 2.482 | 0.403 | 1.689 | 3.276 |

## Discussion

The results of regression analysis showed the similar finding from both Thailand and Indonesia between the relationship of the functional value, emotional value, social value, monetary value and satisfaction.

The functional value affects satisfaction for mobile Internet users both in Indonesia and Thailand. This implies that mobile Internet users realize or well perceived ability of mobile Internet and it affect the user satisfaction at the end.

Emotional value can significantly affect the satisfaction of respondents in Indonesia but not in the case of Thailand. Indonesian respondents shows that there is a tendency to use mobile Internet to meet theirs emotional needs, either by having the device or to access Internet services that they like. Hedonic criteria related to feelings and emotions such as excitement, relaxation or pleasure (Kim et al., 2004). Similarly, Arambewela et al. (2005) found Indonesian respondents have a tendency to greater interest to the values associated with hedonism. This is consistent with the results of multiple regression. However, this factor is insignificantly affecting the satisfaction of Thai mobile Internet user. It may imply that Thai mobile Internet users paid more attention to other factors when using mobile internet.

The result of Indonesian social value shows that significantly affect satisfaction. The test results of Indonesian social value align with the explanation of Subagyo (2009) that the people of Indonesia have high social network characteristics.. In Thailand, social value is also significantly affect satisfaction. This is consistent with Nguyen (2005) which states that the Thai people really appreciate the friendship and tend to look for a permanent friendship. In addition, the mobile Internet services most frequently and widely used by respondents in Thailand in this study are social networking.

In addition, the effect of monetary value on satisfaction is the same for both countries. Results indicate that monetary value can significantly affect satisfaction.

Monetary cost is an important to be considered by consumers of mobile Internet since they are hoping to get the best services in accordance with costs that they have to spend. Monetary value means how satisfactory mobile Internet services compared with the cost, time or effort to be spent in using the mobile Internet (Sweeney and Soutar, 2001).

However, there are the different between each value factors in these two countries according the t-test. The results of these differences are influenced by the perceptions of respondents in Indonesia and Thailand to the different mobile Internet usage pattern. In this case it turns out different cultures influence the way in perceiving the use of different technologies, especially in the mobile Internet usage pattern. The results of this research can provide us more knowledge about cultural differences in the usage patterns of mobile Internet, especially in Indonesia and Thailand. In the results of this study indicate that statistically there are differences in usage patterns of mobile Internet that is influenced by the culture of the measured value.

## Conclusion

This study aims to examine factor influencing mobile Internet usage by including a perspective of intercultural in developing countries. Thailand and Indonesia mobile Internet user were the sample in this study. Data obtained from survey in 2013. The multiple regression analysis and t-test was employed to analyze the relationship between the value structures and overall satisfaction in each country. Results confirm that the value structure significant impact satisfaction of mobile Internet user in both country with different degree.

The results of this study can contribute to the government in order to assist in giving the communications department policies related to tariff policy in order to give more wisely tariff policy and more affordable for the whole society and its stakeholders. Findings also provide information to strategic manager of service providers regarding to the cost and pattern of usage in order to utilize the right strategies for each country. Future research is still needed. The different uses may also be caused by differences in demographics, the infrastructure of the mobile Internet, and mobile Internet services themselves. Thus, examining the relationship between the structure of values and behavior patterns of users could be considered as an additional study.

## References

Arambewela, R., Hall, J., Hede, A.M. (2005). Personal values of international postgraduate students from Asia: A Cross Cultural Study. In ANZMAC, Conference: Marketing Education.

Duncombe, R. (2011). Researching impact of mobile phones for development: concepts, methods and lessons for practice. *Information Technology for Development*, 17 (4), 268–288.

ITU (International Telecommunication Union). (2009). *Measuring the Information Society: The ICT Development Index*, Geneva, Switzerland

ITU (International Telecommunication Union). (2012). *Measuring the Information Society*, Geneva, Switzerland.

Johnson, M.D. & Fornell, C. (1991). A framework for comparing customer satisfaction across individuals and product categories. *Journal of economic psychology*, 12, 267-286.

Kalakota, R. & Robinson, M. (2002). *M-Business: The Race to Mobility.* New York:McGraw-Hill Companies Inc.

Kim, J., Lee, I., Lee, Y. & Choi, B. (2004). Exploring E-business implications of the mobile Internet: A cross-national survey of Hong Kong, Japan and Korea", *International Journal of Mobile Communication,* 2(1), 1-21.

Kreitner, R. & Kinicki, A. (2007). *Organizational Behavior*, New York: McGraw-Hill.

Lee, Y., Kim, J., Lee, I. & Kim, H. (2002). A cross-cultural study on the value structure of mobile Internet usage: Comparison between Korea and Japan. *Journal of Electronic Commerce Research,* 3(4).227-239

Mulder, N. (2000). *Indonesian images, the culture of the public world*. Yogyakarta: Kanisius Publishing House.

Nguyen, T. H. (2005). Thailand: Cultural background for ESL/EFL teachers. Retrieved from http://bit.ly/XIBQwR ( 2 February 2014)

Sheth, J.D., B.I. Newman, & Gross, B.L. (1991). Consumption values and market choice. Cincinnati: South Western Publishing, Cincinnati.

Subagyo, Y. (2009). *Pemberdayaan Masyarakat Melalui Posdaya Untuk Mengatasi Krisis Sosial,* Himpunan Pandu dan Pramuka Wreda (HIPPRADA),

Sweeney, J.C. & Soutar,G.N.(2001). Consumer perceived value: The development of a multiple item scale. *Journal of Retailing*, 77, 203-220.

Wardani, R., Warsono, S. (2012). A Cross-cultural study on the value structure of mobile Internet usage: Comparison between Indonesia and Japan, *Journal of Indonesian Economy and Business,* 27 (3), 355-372.

**Contact email:** kschalit@kmitl.ac.th, chalita.srinuan@gmail.com

# Adoption of Thai rubber Smallholders' household on the use of ICT: A case of Southern Thailand.

Chalita Srinuan, King Mongkut's Institute of Technology Ladkrabang, Thailand
Amnuay Seangnoree, King Mongkut's Institute of Technology Ladkrabang,Thailand

The Asian Conference on Technology, Information & Society 2014
Official Conference Proceedings 2014

**Abstract**
Rubber is one of economic crops which create high economic values for Thailand apart from rice. Under the uncertain socio-economic and environment situation, governments have attempted to overcome some of the perceived information failures related to production via agricultural extension services for example technology transfer program, e-service and e-library. Thailand is still very young in term of Information and Communication Technology (ICT) adoption in agricultural sector, however. This study aims to examine empirically whether socio-economic variables and farmer's perception have a systematic link with the adoption of ICT service by employing the Technology Acceptance Model (TAM). Data were gathered from a distributed questionnaire in the South of Thailand during December -May 2014 made up of 264 small rubber farmers. The findings reveal the socio-economic variables have indirect effects on the ICT adoption via the perceptions on rubber smallholders household. The significant perceptions are perceived of usefulness, perceived quality, social influence and facilitating condition. Possible policies to promote more ICT adoption among rubber smallholders were also discussed.

**Keywords:** ICT adoption, Perceived usefulness, Perceived quality, Rubber smallholders, Thailand

## Introduction

Agriculture can serve as an important engine for economic growth in developing countries. Among economic crops, rubber is one of them which create high economic values for Thailand apart from rice. At the end of 2011, Thailand is the world's largest producer of natural rubber (OAE, 2011) and also the world leader in rubber wood production and export. Moreover the rubber plantation area has been increasing in Thailand and plantations have started to spread from the South to new areas in the East and Northeast of Thailand. The rubber plantation area in Thailand is much larger than the area of forest plantations in the country. According to OAE (2011), the total area of rubber plantations in Thailand was 3,001,760 hectares (ha) in 2011. Majority of rubber farmer is a smallholder and has small rubber plantation which is 95% of total rubber plantation in the country (RRIT, 2010). The Department of Agriculture (DOA) of Thailand has classified smallholdings, medium-sized holdings and estates as those where rubber area is less than 8 ha, 8-40 ha and more than 40 ha, respectively (Pratummintra 2005). Specifically, the smallholding is usually family-owned, managed by the family head and worked by family labour, whereas the medium-sized holding and estates are frequently owned by a company or a government enterprise, and usually professionally managed.

In recent years although latex is still the main product of rubber plantation, wood selling can increase the total productivity and enable reaching a maximum productivity of the rubber plantation earlier. Regarding productivity in the wood industry, there has been rapid growth in the value of rubber wood, which has generated into an industry in its own right (Watcharakuldilok & Vitayaudom, 2008). However, under the uncertain socio-economic and environment situation such as rubber price and climatic variation are effect on rubber product and production change of rubber smallholders household that goal for increase their income and also, negative effect on farm management as labor constraint in both quality and quantity high cost of input factor. The rubber smallholder is intensively supported by the government, in forms of technology and production inputs such as seedlings, land preparation and fertilizer. Moreover, the governments have attempted to overcome some of the perceived information failures related to adoption via agricultural extension services, i.e. technology transfer program, e-service, e-library and etc. Thailand is still very young in term of ICT adoption in agricultural sector, however.

Bailee, Chockchaisawadee, Putthongsri, Jitthaisong and Bailee (2008) showed that the most common form of ICT use by a Thai rubber planter was to access news and information from traditional media such as television and radio. Nikomborirak and Pongsukcharoenkul (2011) also confirmed that productivity of the agricultural sector is low due to four main factors including deteriorating soil quality, an insufficient water source, a suboptimal scale of operation and logistics cost. In each product supply chain case, ICT is applied to enhance the productivity or cut production costs borne by farmers. The water management case does not focus on the use of ICT in any specific agricultural supply chain but rather, illustrates the use of ICT to improve the management of agricultural production.

Considering ICT using by region in 2012, the proportion of using computer, Internet and mobile phone, Bangkok had the highest proportion of computer users at 51.5% and next was the Central region with 35.2%, Southern region with 32.0%, Northern region with 31.6% and Northeastern region had the lowest with 29.1%. For the using

of Internet, Bangkok also had the highest proportion of users at 44.4% and the second region was the Central region at 27.5 Southern region at 25.5%, the Northern region was 25.0% and the Northeastern region had the lowest proportion at 21.5 %( NSO, 2012). NSO reported also indicated that there are gaps between people who work in agriculture industry and other industries in term of using computer and internet usage. Considering possession of mobile phone, Bangkok had the higher percentage of mobile phone owner (80.0%) while Southern and Northern region had 67.8% and 64.1% respectively.

Another important step is education in the use of ICT and according to Nikomborirak and Pongsukcharoenkul (2011), almost 85% of Thai farmers have only completed primary education. Which such little education, Thai farmers are easily exploited by the middleman or other parties in the supply chain. For example, rubber is sold in multiple forms and the determination of its prices depends upon a multitude of factors. The information on prices is often passed down the value chain and cheating is easy during this process but according to De Silva and Ratnadiwakara (2008), mobile telephony can really help make farmers better off.

Hence, this study aims to examine empirically whether socio-economic variables and farmer's perception have a systematic link with the adoption of ICT services, in particular mobile phone, by employing the Technology Acceptance Model (TAM) model. The findings should be of interest both to academics and practitioners. From a theoretical perspective, this study provides a model for identifying antecedents of user intentions as a contribution toward the larger effort to adopt ICT in agricultural sector. From a practical standpoint, the findings should guide an industry that is promoting the ICT as a tool to attract farmers by enhancing usability and accessibility, as well as ensuring quality. The findings will offer insights into the implications of farmer's adoption of the ICT as a tool in their job. Government and policy makers are facing opportunities as well as challenges in developing new forms of ICT in order to promote more ICT adoption in agricultural sector.

**Prior studies**

There are various competing theories for technologies acceptance in extent literature that are used by information system researchers to assess the user's intention to adopt new technology. Among different models that have been proposed, Giovannis et al. (2012) noted that the technology acceptance model (TAM) suggested by Davis (1989) is the most widely used model for three reasons. First, it is focused specifically on information system use; second, it is based on social psychology theory; and third, it shows parsimony and empirical support from various studies. TAM is adaptation of the theory of reasoned action (TRA) by Fishbien and Ajzen (1975). The key purpose of TAM is to provide a basis for tracing the impact of external variables on internal beliefs, attitudes and intentions. It suggests that perceived ease of use (PEU), and perceived usefulness (PU) are the two most important factors in explaining system use (Legris, Ingham & Collerette, 2003).

Flett et al. (2004) sought to determine whether the TAM could adequately explain adoption and use of dairy farming technologies. Their finding revealed that perceived usefulness (PU) and perceived ease of use (PEU) are significantly greater for farmers using technologies compared to those that are not. Similarly, De Silva et al. (2011) showed that that people at the Bottom of the Pyramid (BOP) are intend to disconnect

if they do see as much economic benefit arising from access to telecom. There are many costs and burdens for the BOP in the agricultural sector. De Silva and Ratnadiwakara (2008) also found that the total transaction costs are 15 percent of total cost that smallholder vegetable farmers in Sri Lanka confront with. They showed that if the farmers use ICT by simple phone call, the transaction cost would reduce substantially. Pick et al. (2013) employed TAM framework to understand factors that influence adoption and use of information technology by farmers. Their results support the dimensions suggested by Rogers: relative advantage, compatibility, low complexity, and observability, while also supporting the TAM factors of usefulness and ease of use. However, Wijerthna (2011) noted that despite the farmer had access and familiar with the telephone, their level of awareness on various existing telephone based agrarian service was low.

Venkatesh et al. (2003) developed the Unified Theory of Acceptance of use of technology (UTAUT) model to explain user intention to use an information system and subsequent usage behavior. Regarding to UTAUT, performance expectancy, effort expectancy, social influence and facilitating conditions are the key determinant factor of usage intention and adoption. At the same time demographic factor (age, gender, experience and voluntariness of use) are mediating factors in the impact of usage intention and adoption. van Biljoin and Kotzé (2008) employed the UTAUT model to examine the mobile adoption. They concluded that mobile adoption is influence by demographic, social, cultural and contextual factors.

Zhou (2012) utilized the UTAUT and privacy risk, to examine user adoption of location-based services. The results indicated that usage intention is affected by both enablers such as performance expectancy and inhibitors such as perceived risk. Lai and Lai (2014) employed a technology acceptance model for m-commerce with five factors is constructed. The acceptance of m-commerce is influenced by factors including performance expectancy, social influence, facilitating conditions and privacy concern; while effort expectancy is insignificant in this case. In additional, Touray et al. (2014) identified relevant elements of Internet adoption at the user level in The Gambia through UTAUT. Their results suggested that Internet adoption at the user level in The Gambia can be viewed as a three-layered pyramid. It consists of seven moderating factors (age, gender, experience, voluntary use, friends' influence, Internet service providers and regulators), four indirect determinants (performance expectancy, effort expectancy, social influence (SI) and facilitating conditions (FC) and three direct determinants (education, behavioral intention and income).

Moreover, infrastructural support or perceived of quality (PEQ) is also another important factor which confirmed by the study of Shin (2012). The intention to accept a new technology is determined by users' perception on relevant quality variables (call, service, mobility and coverage). Thus, the perceived of quality will also be examine in this study.

In sum, the prior studies show that TAM and UTAUT provide an extremely useful theoretical tool for understanding how peoples' technology acceptance level impacts their intention to use. There have been various model developed for integrating TAM. TAM has gained popularity across sectors, but there is a gap for applying TAM in agriculture sector. This study will discover the adoption of ICT, in particular mobile telephony, by investigating ICT usage among rubber smallholders as a case.

From the theoretical and empirical background, the research model, that aims to examine empirically whether socio-economic variables and perceptions have a systematic link with the decision of rubber smallholders household on ICT adoption by integrating the TAM model and UTAUT model is illustrated in Figure 1.

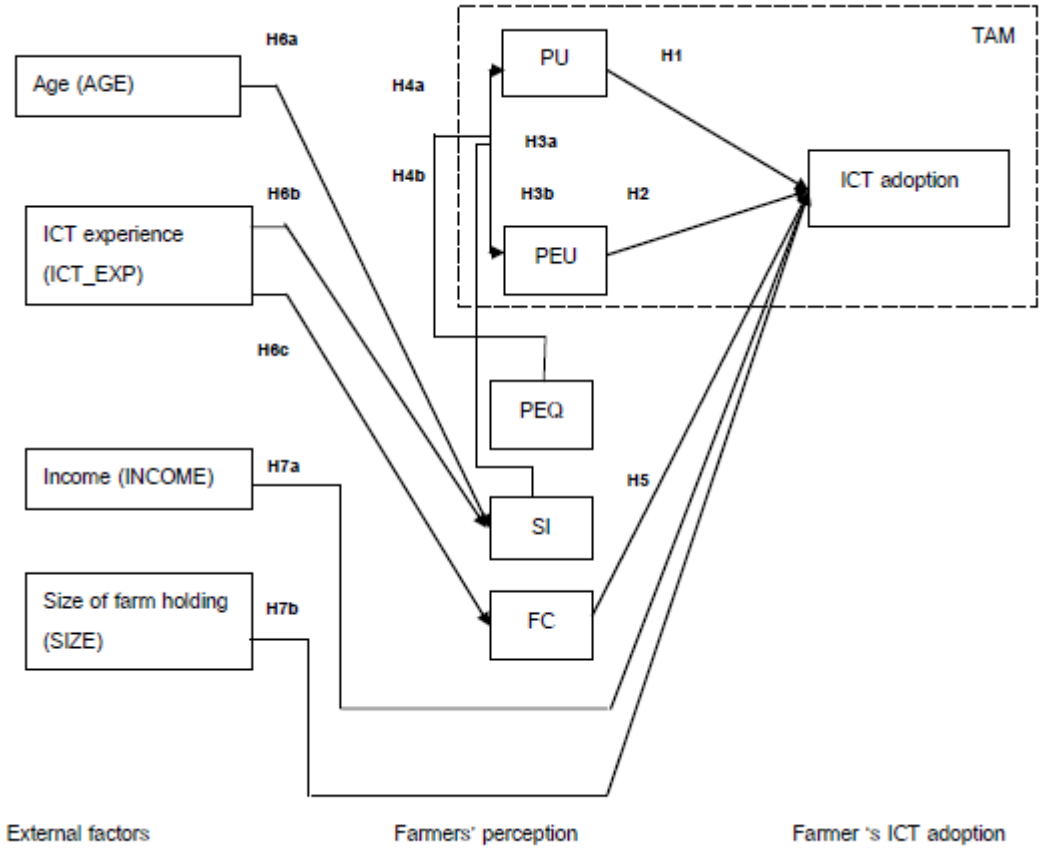



Figure 1: Framework of the study.

**Data and method**

This study employed the quantitative researcg method. Data was collected from the rubber smallholder's household in three province provinces (SuratThani, Songkhla and Nakorn Si Thammarat) in the South of Thailand. The total distributed questionnaire is 435 samples. A survey was conducted through a semi-structured interview. During the survey questionnaire was comprised of three components, which are a) socio-economic factor, b) experience in using ICT and c) ICT perception and acceptance of rubber farmers.

The Likert scales (1-5), with anchors ranging from "Strong disagree" to "Strongly agree" were used for all belief items to ensure statistical variability among survey responses for all items measured. A questionnaire was designed, measured the reliability and consistency, and distributed to collect data from the small rubber farmers in the selected provinces in the South of Thailand through a semi-structured interview. Finally, 264 usable, complete responses were obtained or 61.0 percent for response rate.

## Findings

The survey results show that there are three main services available for respondents and the main ICT service that rubber smallholder's household access most is the mobile phone (98.1.0 percent) and followed by Internet (14.0 percent) and fixed telephony (11.4 percent) respectively. Respondent has a same degree of usage experience both fixed telephony (6.8 months) and mobile telephony (6.9 months), while the experience on Internet usage (4.5 months) is less than those two services. The overall experience of ICT usage among small rubber farmers is 7.9 months.

Overall of adoption on ICT attributes is low. Findings present that rubber farmer rarely uses the ICT services related to their farming activities. Interestingly, the attributes that rubber farmer uses most on average are exchanging information with other farmers and followed by searching the price of product from a middleman and following product price. At the same time, they never or rarely use attribute on getting and searching news on agriculture technology from government agency and followed by contacting and exchanging information with the government office and following input price.

This study employs a path analysis with the framework of TAM to develop a model that represents the relationship of external factor and farmers' perception toward ICT adoption among the rubber smallholder households. The final fit model is presented as shown in Figure 2. Figure 2 shows that the external factors which have influence on the ICT adoption are age, ICT usage experience and farm size. These factors reveal indirect influence the ICT adoption through the perceptions on rubber smallholders household. The significant perceptions are PU, PEQ, SI and FC.

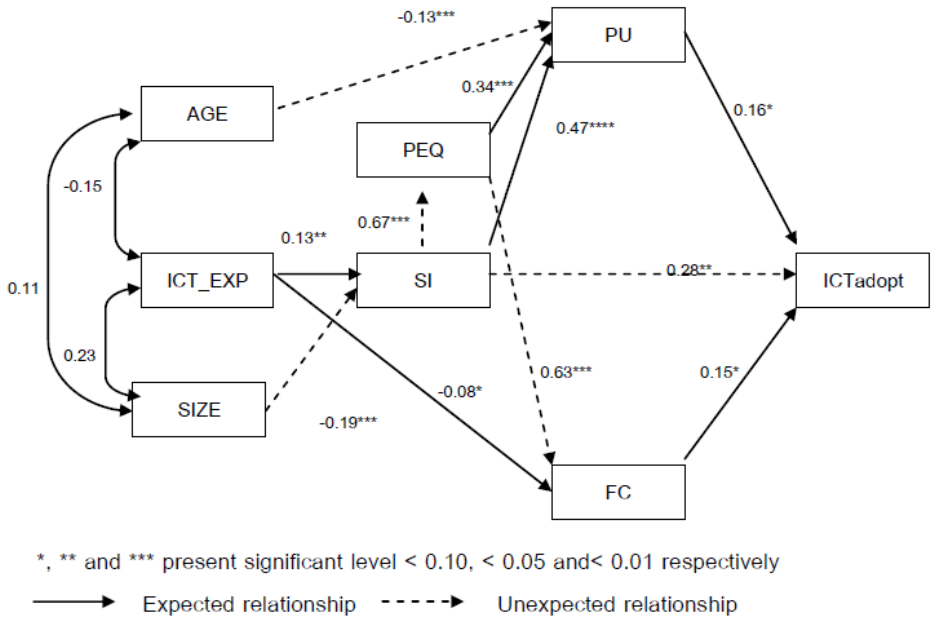



Figure 2: Findings of the study

## Discussion

The results of this study consistent with previous studies on technology adoption, the findings show that perceived usefulness (PU) has a considerable impact on ICT

adoption (Flett et al., 2004; Aubert et al., 2012). In particular, it is demonstrated that this variable is greatly influenced by age of rubber farmer, PEQ and SI as confirmed by Shin (2012).These reflect that the effect of PU and perceived quality (PEQ) and social influence (SI) are found to be valid and significant in the study. In addition, facilitating conditions (FCs) also reveals its role on ICT adoption in the study which is similar to the study by Jain and Hundal (2007) and Seneler et al. (2008). The result warrants a consistent model for the drivers of ICT adoption among rubber farmers.

Some of the pre-setting hypotheses do not show the relationship as expected, for example, the relationship between PEU and ICT adoption, PEQ and PEU, and SI and PEU. This may due to the ICT service that majority of rubber farmer use is the mobile telephony. The mobile service is not complicated to use in terms of skilled needed as compared to Internet. Rubber farmers can learn by himself or consult with their friends. Moreover, income and size of farming holding did not have influence on ICT adoption as the literature suggested. One possible reason could be the cost of using mobile phone is not expensive. Age and income may not a main obstacle to adopt the service. While the size of farm holdings has no relationship directly with ICT adoption as earlier research established (Yengoh et al, 2009; Poolsawas & Napasintuwong, 2013; Kakuru et al., 2014), but it has indirect influence through other variables.

Interestingly, there are unexpected relationships from the path analysis. Firstly, the relationship between AGE and PU which shows that age has a negative direct effect on perceived usefulness. This refers that the younger rubber farmer, the greater perceived usefulness of ICT service. This will also impact the ICT adoption. Secondly, size of farm holding has negative influence on social influence (SI). This may imply that rubber farmer who owns a smaller size of the rubber farmers need more supports from their society, including friend and colleagues. Thirdly, the impact of SI on perceived of quality (PEQ) shows positively significant relationship. This suggested that PEQ, i.e. call, service, mobility and coverage, is determined by social influence of user. Friends or colleagues can provide a suggestion on the ICT service that suite farmer needs. Fourthly, PEQ has the positively significant impact on facilitating conditions (FC). This may indicates that farmer still needs the facilitating means in order to adopt ICT service though the quality of service is good. Lastly, SI has the positively significant effect on ICT adoption. This suggests that the ICT adoption of the rubber farmer depend on friend and colleagues. It may reflect the network effect phenomena. If farmer uses the service under the same service provider, the more valuable the service is to each owner.

Considering the significant variables, the role of policy makers and regulator are needed. The government together with related units needs to encourage and support for ICT usage among farmers. This support will make rubber farmers realize that using ICT service will help them to gain higher performance if they use the useful related applications. Availability of ICT service through the public community shared can be another alternative. In terms of accessibility, farmer will adopt more ICT service if the facilitating environment is well-provided. Training course on ICT knowledge and technology could be one of the examples. In addition, Elderly farmer needs special assistance on using ICT. Hence, the facilitating conditions are important in this case. At the same time, regulator also plays an important role to guarantee the quality of service and to protect consumer interests through promoting fair, efficient

and sustainable network competition. Findings confirm that quality of service is a crucial factor for adoption, though it has an indirect effect on ICT adoption in this case. Regulator should be well-prepared to address and consider the following issues; signaling network coverage, changing of price plan without notification, Internet speed guarantee. This will ensure that ICT adoption is beneficial to rubber farmer**.**

**Conclusion**

This study aims to examine empirically whether demographic variables and farmers' perception has a systematic link with the decision of rubber smallholders household on ICT service adoption by integrating the TAM model and UTAUT model. The questionnaires are distributed to collect data from the small rubber farmers in the selected provinces in the South of Thailand through a semi-structured interview during December 2013-May 2014.Completed questionnaire were 264 or 61.0 percent for response rate. Path analysis are utilized for hypotheses testing.

Findings show that the external factors which have influence on the ICT adoption are age, ICT usage experience and farm size. These factors reveal indirect influence the ICT adoption through the perceptions on rubber smallholders household. The significant perceptions are perceived of usefulness, social influence, facilitating condition and perceived of quality. These factors affect the ICT adoption among rubber farmers. Government need to consider the possible policies to promote more ICT adoption among rubber smallholders.

Future research is also needed to fulfill the knowledge in this study. Availability of longitudinal data is necessary and will be benefit or future research. It offers an opportunity to gain a deeper understanding. This study provides the overview picture of ICT adoption, which mainly represents mobile adoption. Future study could elaborate more on this issue by focusing on specific service such as mobile Internet. The mobile Internet can be considered as an emerging technology which may suit with the lifestyle of rubber the farmers rather than other technologies.

## References

Aubert, B.A., Schroeder, A. and Grimaudo, J. (2012). IT as enabler of sustainable farming: An empirical analysis of farmers' adoption decision of precision agriculture technology. *Decision Support Systems, 54*(1), 510-520.

Bailee, S., Chockchaisawadee, S., Putthongsri, P., Jitthaisong, O and Bailee, C. (2008)**.** A study of potential and model of para rubber planters development. Research paper, National Council research of Thailand and Loei Rajabhat University**.**

Davis, F.D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of Information technology. *MIS Quarterly 13* (3), 319-340.

De Silva, H. and Ratnadiwakara, D (2008). Using ICT to reduce transaction cost in agriculture through better communication: A case-study from Sri-Lanka. Avaialble at http://bit.ly/1uDaVOI (Accessed on 1 May, 2014)

De Silva, H., Ratnadiwakara, D. and Zainudeen, A. (2011) Social influence in mobile phone Adoption:Evidence from the bottom of the pyramid in Emerging Asia*. Information technologies & International Development, 7*(3), 1-18.

Flett, R., Alpass, F., Humphries, S., Massey, C., Morriss, S., & Long, N. (2004).The technology acceptance model and use of technology in New Zealand dairy farming. *Agricultural Systems, 80*(2), 199–211.

Fishbien, M. and Ajzen,I. (1975) . *Belief, Intention and behavior: an introduction to theory and research*. Addison Wesley.

Jain, A., and Hundal, B.S.(2007). Factors influencing mobile services adoption in rural India. *Asia Pacific Journal of Rural Development*, 17(1), 17-28.

Kakuru, O.V., Doreen, M. and Wilson, M.(2014) Adoption of On-Farm Tree Planting in Kibaale District, Western Uganda. *Journal of Sustainable Forestry*, 33 (1), 87-98.

Legris,P., J. Ingham & P. Collerette (2003). Why do people use information technology? A critical review of the technology acceptance model.

National Statistical Office (NSO). (2011). *The 2011 Establishment Survey on Use of Information and Communication Technology*. Bangkok: Thailand.

Nikomborirak, D. and Pongsukcharoenkul, N (2011). Towards a knowledge based economy: Case studies of the use of ICT in enhancing agricultural value chains in Thailand. Available at http://bit.ly/1APkGdG (Accessed on 9 May, 2013)

Office of Agricultural Economics (OAE) (2011). Agricultural statistics of Thailand year book 2011. Available at http://bit.ly/1oNPbLm (Accessed on 11 September, 2012)

Poolsawas, S. and Napasintuwong, O.(2013). Farmer innovativeness and hybrid maize diffusion in Thailand. *Journal of International Agricultural and Extension Education*, 20 (2), 50-63.

Rubber Research Institute of Thailand (RRIT) (2010). Rubber Academic Data (in Thai). Available at http://bit.ly/1q67Dhu (Accessed on 17 September, 2012).

Seneler, C.O., Basoglu, N. and T.U. Daim. (2008). A taxonomy for technology adoption: A human Computer Interaction Perspective. PICMET Proceedings, South Africa.

Shin, D-H. (2012). What make consumers use VoIP over mobile phone? Free riding or consumerization of new service. *Telecommunications Policy, 36*, 311-323.

Touray, A. , Salminen, A. , Mursu, A.b (2014). Internet Adoption at the User Level: Empirical Evidence from The Gambia. *Information Technology for Development*, Article in Press.

van Biljoin, J. and Kotzé, P. (2008). Cultural factors in a mobile phone adoption and usage model. Journal of Universal Computer Science, 14(16), 2650-2679.

Ventakesh , V., Morris, M.G., Davis, G.B. and F.D., David (2003) User acceptance of information: Toward a unified view, MIS Quarterly 27(3), 425-478.

Yengoh, G.T, Ato, A.F. and Svensson, M. (2009). Technology adoption in small-scale agriculture: The case of Cameroon and Ghana. *Science, technology & Innovation Studies*, 5(2), 111-131.

Zhou, T. (2012). Examining location-based services usage from the perspectives of unified theory of acceptance and use of technology and privacy risk, *Journal of Electronic Commerce Research*, 13 (2), 135-144.

**Contact email:** kschalit@kmitl.ac.th, chalita.srinuan@gmail.com

# *Performance Evaluation of TCP/IP vs. OpenFlow in INET Framework Using OMNeT++, and Implementation of Intelligent Computational Model to Provide Autonomous Behaviour*

Pakawat Pupatwibul, University of Technology Sydney, Australia
Ameen Banjar, University of Technology Sydney, Australia
Robin Braun, University of Technology Sydney, Australia

The Asian Conference on Technology, Information and Society
Official Conference Proceedings 2014

**Abstract**
Analysing performance of transmitting data from a source to a certain destination is an interesting task. One of the most reliable networking protocol suites is the Transport Control Protocol and the Internet Protocol (TCP/IP), which will be studied against a new management paradigm called Software Defined Networking (SDN). SDN is an emerging programmable network architecture, where network control plane is decoupled from forwarding plane. SDN forwarding methods are based on flows, which operate in contrast to conventional routing methods, such as TCP/IP routing table and MAC learning table. Moreover, OpenFlow protocol has efficient forwarding methods to push L2-L4 functions which are simplified into a Flow-Table(s) abstraction. This paper discusses the relationship between the processes of forwarding packets in conventional IP routing table vs. OpenFlow-table and evaluates the performance between both implementations using INET framework in OMNeT++. While TCP performs slightly better than OpenFlow with respect to mean round trip time (RTT). The results also proved the correctness of OpenFlow implemented simulation model. Finally, we propose a Distributed Active Information Model (DAIM) within OpenFlow to support an autonomic network management.

Keywords: OpenFlow, Network Performance, TCP/IP, Software- Defined Networks, OMNeT++.

## Introduction

In the last few years network technologies have been improved significantly in performance, complexity, functionality and other aspects, because of current needs and necessities of the modern world. The Internet protocol suite, widely known as TCP/IP, is a networking model and the basic communication language or protocols used to connect hosts on the internet. TCP/IP is the best known protocol suits today because of the successful development of the internet, and thus useful to study the behaviours of this protocol further, by making use of simulations.

The Transmission Control Protocol (TCP) and the Internet Protocol (IP), commonly known as the TCP/IP standard, is widely used for network communications. TCP/IP attempts to make efficient use of the underlying network resources, by specifying how data should be transmitted, formatted, addressed, routed and received at the destination (Forouzan, 2002). TCP/IP was developed to support maximum throughput over many kinds of different networks. Although TCP/IP supports many current network services, it is not efficient for the requirements of business needs and end users. This has led to the development of alternative networking architectures, and the introduction of Software-Defined Networking (SDN).

However, OpenFlow is still not widely standardised yet because each year Open Network Foundation (ONF) has introduced newer versions including improved and extra functionalities. The latest version is known as OpenFlow specification 1.4 (ONF, 2012). The evolution of OpenFlow protocol versions are kept experimenting within labs. This chapter compares the performance of TCP/IP with the newly emerging OpenFlow standard for software defined networking.

A variety SDN approaches have been developed, but there is only limited information on the performance of each, and realistic performance comparisons are not widely available. Simulation tools such as the OMNeT++ INET Framework are suitable for the task of designing, building, and testing network architectures, and provide practical feedback when developing real world systems (Varga, 2001). Such simulation tools allow system designers to determine the correctness and efficiency of a design before a system is deployed. Simulators also enable the evaluation of the effects of various network metrics, and provide mechanisms to obtain results that are not experimentally measurable on larger geographically distributed architectures. However, very few performance evaluations of OpenFlow architectures using simulation tools such as OMNeT++ INET Framework have been published (Varga, 2010 & Varga, 2012).

The performance comparison simulates OpenFlow and TCP/IP networks using the OMNeT++ INET Framework discrete events network simulator. By analyzing key network metrics including round-trip-time (RTT) and data transfer rate (DTR), the results indicate that OpenFlow performed slightly better than TCP/IP in this analysis. The results also proved the correctness of OpenFlow implemented simulation model.

Thus, this paper evaluates the performance of TCP and OpenFlow implementation in INET framework 2.0 using OMNeT++. In addition, the paper covers a number of multiple running from the simulation to preserve accuracy of the simulation results. The remainder of this paper is organized as follows. Sect. 2 describes the motivations of the study. In Sec. 3, we introduced the background and overview of TCP/IP and

OpenFlow networks. We present more details of related works on network simulation and emulation tools in Sect. 4. In Sect. 5, the performance of TCP/IP vs. OpenFlow is evaluated in the simulations. Section 6 proposes the DAIM model within OpenFlow to support an autonomic network management. Finally, Sect. 7 concludes this work.

**Motivations**

OMNeT++ can be applied to different network scenarios, topologies, behaviours and environments or other application fields for different purposes. Then, it can be studied and analysed to see how the system is functioning and performing. For example, applications of networking simulation area include network traffic, data transfer rate, packet counts and round trip time for packets transmission. OMNeT++ will be the first step for Australia when attempting to implement a new network infrastructure such as OpenFlow. OMNeT ++ is easy to simulate geographic distance and help predict how that would affect the behaviours of this new infrastructure, when considering different technologies or products running on different software. Thus, we have used OMNeT++ modules to design, simulate, verify, and analyse the performance of different networks protocols, where in this context we used TCP/IP and OpenFlow.

OpenFlow can offer network administrators the flexibility to manage, configure and optimise network resources, and thus can effectively control the network behaviour in real-time as well as deploying new network applications and services. OpenFlow-Based SDN can present several substantial benefits including centralised management and control of network devices from various vendors, the direct manipulation of new network services without having to configure each node individually, programmability by administrators, enterprises, users, and software vendor, and the ability to provide centralised and automated management which increases network security and reliability.

Currently, a related work of integrating the OpenFlow protocol version 1.2 in the INET framework for OMNeT++ has been developed. The motivation is to test the correctness of their implemented model in overall compared to TCP modules in INET, and focus especially on the performance of controller's placement based on a variety of performance metrics in the investigated network.

**Background of TCP/IP and OpenFlow network**

This section introduces Transmission Control Protocol/Internet Protocol (TCP/IP) and provides a quick introduction to OpenFlow-Based SDN, discussing what they are all about in an overall context. We begin by defining both TCP/IP and OpenFlow network in the most general terms.

**A. Overview of TCP/IP**

The Transmission Control Protocol and Internet Protocol (TCP/IP) is a suite of communication protocols widely used to connect hosts on the Internet and on most other network communications as well (Fall et al., 1996). TCP operates at the transport layer, the middle layer in the seven layers of the OSI model (Jeroen et al., 2004). This layer maintains reliable end-to-end communications between network devices. On the other hand, IP is a network layer protocol, which is responsible for packet forwarding including routing across intermediate routers (see Table 1.).

Because TCP/IP was developed earlier than the OSI 7-layer mode, it does not have 7 layers but only 4 layers.

<table>
<tr><th>TCP/IP Protocol Suite</th><th>TCP/IP</th><th>SDN</th></tr>
<tr><td>FTP, SMTP, Telnet, HTTP, ……</td><td>Application</td><td>Application</td></tr>
<tr><td>TCP, UDP</td><td>Transport</td><td rowspan="3">Control Layer</td></tr>
<tr><td>IP, ARP, ICMP</td><td>Internet</td></tr>
<tr><td rowspan="2">Network Interface</td><td rowspan="2">Network Access</td></tr>
<tr><td>Physical</td></tr>
</table>

Table 1: Comparison of OSI, TCP/IP and OpenFlow models

One fundamental feature of the IP protocol is that it only deals with packets, addresses, and directing messages to where they are intended (Stewart et al., 2001). This is the most significant unit of TCP/IP data transmission. TCP allows two devices to complete a connection and exchange streams of data. TCP assures that the data and packets will be delivered to the destination in the same order in which they were sent.

Like the OSI model, functionalities of TCP/IP has been organised into four abstraction layers. (1): Network Access layer contains the network interface card which provides access to the physical devices. (2): Internet layer establishes network to network communications and therefore connects internetworking. (3): Transport layer handles the end-to-end (host-to-host) communication. (4): Application layer offers the users with the interface to communication and gives a way for applications to have access to networked services.

TCP/IP is a set of protocols developed to allow cooperating computers to share resources across a network and also to ensure network's robustness by recovering automatically from any failure of nodes on the network. Furthermore, it can allow large scaled networks to be contracted with minimal requirements of central management.

**B. Overview of OpenFlow-based SDN**

Software-defined networking (SDN) is a relatively advanced method for implementing communication networks (McKeown et al., 2008). SDN separates the decision maker, called the control plane, which decides where packets are sent, from the underlying infrastructure, called the data plane, which forwards packets to the decided destination. This is a migration of control can formerly and tightly bound in individual network devices enabling the underlying infrastructure to be abstracted for applications and network services, which can treat the network as a logical or virtual entity. A newly emerging standard for SDN is the OpenFlow standard, which includes a standardized protocol for communications between the control plane and the data plane (ONF White Paper, 2012).

OpenFlow was initially introduced by Stanford University in 2008, as the first standardised communication interface defined between the control plane and the data plane of the SDN architecture (ONF, 2012). OpenFlow is an open standard that enables researchers to run experimental protocols in the networks without having to expose vendors' internal implementations of their network devices. In classical switches and routers, the fast packet forwarding (data plane) and the high-level routing decision (control plane) happen in the same network element. In OpenFlow, it separates these two functions. The data plane portion still resides in the switches, whereas the routing decisions are moved to a different device called the controller, typically a standard server. Figure 1 shows the communication between controller and OpenFlow switch trough a Transport Layer Security (TLS) and its predecessor, Secure Sockets Layer (SSL) channel using the OpenFlow protocol (OpenFlow Switch Consortium, 2009).

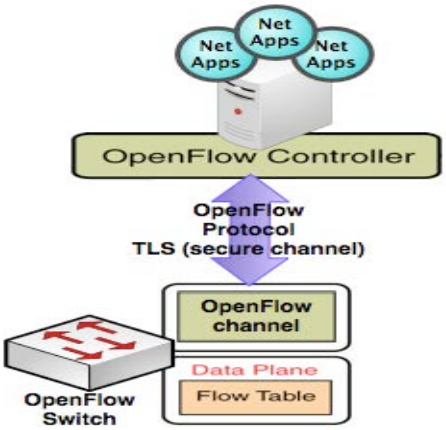



Figure 1: OpenFlow-Based SDN structure

**Related works**

This paper focuses on evaluating and analyzing networks performance and results, in this regards we use OMNeT++ network simulator. OMNeT++ has many advantages as a network simulator it's been used widely in research academia fields because; it is able to simulate manufacturing, crowd management, airports and weather forecasting. OMNeT++ is scalable as its all simulated modules are implemented as C++ modules and they are linked together as a single process. Moreover, OMNeT++ can modify parameters such as link bandwidth and delay also it is possible to modify configuration of network size, mobility pattern or speed for performance results corrections (Varga & Hornig, 2008). When time is concerned in OMNeT++ the performance results need to be repeated for correction and accuracy. OMNeT++ supports OpenFlow network as an extension of INET framework including spanning tree protocol (STP).

On the other hand, OMNeT++ has many limitations, one of those limitations relate to OpenFlow protocol is that, it is older than the latest version of OpenFlow of this writing is version 1.4 but this paper uses OpenFlow switch specification 1.2. In addition, OMNeT++ uses C++ modules with its simulation engine code as well as all devices and objects as user-level executable program exactly as ns-3 network simulator (NS-3, 2012). Meanwhile, NOX OpenFlow controller operations is a user-level program but, OMNeT++ and NOX cannot be compiled and linked together to

form a single executable program (Gude et al., 2008). For example, it is compulsory to create C++ module from scratch to build OpenFlow switch or controller, even ns-3 needs to build new modules for the same reason. Therefore, the re-implemented modules may not be the same as the behavior of real device/object with real applications, because the re-implemented module is a much-simplified abstraction of the real devices/objects. As OMNeT++ is supporting many functions such as STP, where STP is not supported by ns-3 as well as no TCP connection between simulated hosts, so in real model when use a TCP connection the results are conceders the packet losses and congestion.

There are many other test-beds for observing the network communication such as EstiNet which combines between network simulator and emulator (Wang et al., 2013). However, emulator has limitations as it is only designed for real-time network functional testing. Whereas the simulation is for arbitrary scenarios, the feature that emulator cannot do, so it does not scale very well. The simulated module of OMNeT++ does not connect to real OpenFlow controller as an external entity for measuring the effects of the OpenFlow protocol where Mininet and EstiNet can do. The network emulator separates namespaces such as Mininet, which lead to reducing the overhead of the system rather being as one simulation process. Mininet is emulated hosts as virtualization approach. Mininet emulated hosts can run real application and can exchange information (Lantz et al., 2010). For example, one of these hosts can be a controller because of the controller is as a real application, also can be an OpenFlow switch or just a normal host which can link to other emulated devices using Ethernet pair (linux kernel).

Network emulator has unpredictable results and different experimental results on each run because of the emulated hosts run based on CPU speed, current system activities, system load, memory bandwidth, number of emulated hosts and multiplexing over the CPU. For example, Mininet schedule packets promptly by the operating system, then it is not guarantee that all software OpenFlow switches will forward packets at the same rate of emulated hosts.

Mininet needs to run up a shell process to emulate each hosts and run up a user-space or kernel-space (OpenVswitch) to simulate each OpenFlow switch. Therefore Mininet is less scalable than EstiNet, ns-3 and OMNeT++. Mininet can only be used to study the behavior of the emulated hosts but cannot be used to study time of network/application performance. Mininet GUI can be used for observation purposes such as observing the packet playback of a simulation run and user needs to write Python code to set up and configure the simulation case. However, OMNeT++ has a GUI, which can be used for result observation and users need to write C++ codes to setup and configure the simulations. Overall, it is better to use OMNeT++ even though it takes more time and effort to create simulations, however once modules are created, it's much easier to create new ones.

**Performance comparison results**

This section provides a performance comparison of TCP/IP vs. OpenFlow modules in INET Framework 2.0 using OMNeT++ network simulation. The network simulator makes it possible to evaluate how a network performs under circumstances with different versions of the protocol stack, and enables analysis of the effects of different variables including channel speed and delay.

### A. Measurement methodology

The OMNeT++ INET Framework 2.0 network simulator is a C++ discrete event simulator. An advantage of this simulator is that it simplifies the integration of new modules, and allows existing modules to be customized. The INET Framework is a network simulation package that contains models for wired and wireless networking protocols, including UDP, TCP, SCTP, IP, IPv6, and Ethernet. The INET Framework has recently implemented an extension to enable OpenFlow to be modeled. The OpenFlow extension is still in early development, and is currently based on Switch Specification Version 1.2 (Klein & Jarschel, 2013).

The OMNeT++ network simulator is used here to simulate the operation of OpenFlow and TCP/IP while logging performance metrics including Data Transmission Rate (DTR) and the mean round-trip-time (RTT) for nodes in the simulated networks. We analyse the DTR and RTT for the nodes within TCP/IP and OpenFlow networks, using a similar network topology for each. Each network includes a number of hosts, two switches and a destination server (see Figure 2). The OpenFlow network also includes an additional device called the controller (standard server), which is directly connected via separate links to the OpenFlow switches. Hence the OpenFlow switches can perform Layer 2, 3, and 4 routing, as compared to the Layer 2 MAC learning table used by TCP/IP. The simulations are logged, and the logs are subsequently analysed to enable the performance of the OpenFlow and TCP/IP networks to be compared (Banjar et al., 2014a).

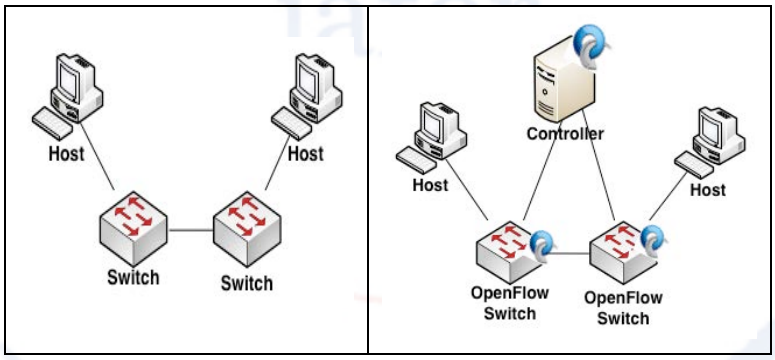



Figure 2: Layout of simulated networks

Performance is compared by varying the traffic and link delays by the same amounts in both the OpenFlow and the TCP/IP simulation channels. In this experiment, we measure the DTR and the mean values of measured RTT between the time-triggered nodes, both for the standard TCP/IP in INET, and OpenFlow. Sequential ping requests are generated, where each includes a sequence number, and replies are expected to arrive in the same order. After the simulation has run for 300 seconds, we measure the RTT for each ping and reply. This results in 48 measurements each for OpenFlow and TCP/IP, for each 300s simulation period. Each simulation was run ten times (OpenFlow and TCP/IP five times each) to reduce simulation artifacts.

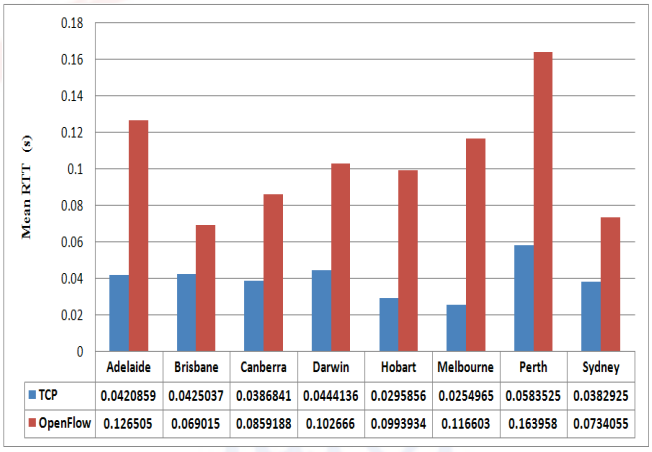
A large number of samples are recorded, and the means and standard deviations of DTR and RTT are computed. RTT is approximated using following equations (Sünnen D., 2011):

$$Tr = \frac{packet\ size}{link\ bandwidth} = \frac{1500 * 8bit}{100 * 10^6 bit/s} = 0.12\text{ms} \quad (1)$$

$$RTT = \alpha * RTT + (1 - \alpha) * Tr \quad (2)$$

The packet size is known to be 1,500 bytes long. The link speed limited to 100 Mbit/s in both the TCP/IP and the OpenFlow networks. ***Tr*** is the transmission time between the segments sent and the acknowledgement arrival, and α is a smoothing factor, which equals the value (7/8) ≈ 0.875.

## B. Result evaluation




| | Adelaide | Brisbane | Canberra | Darwin | Hobart | Melbourne | Perth | Sydney |
|---|---|---|---|---|---|---|---|---|
| TCP | 0.0420859 | 0.0425037 | 0.0386841 | 0.0444136 | 0.0295856 | 0.0254965 | 0.0583525 | 0.0382925 |
| OpenFlow | 0.126505 | 0.069015 | 0.0859188 | 0.102666 | 0.0993934 | 0.116603 | 0.163958 | 0.0734055 |

Figure 3: Mean RTT TCP vs. OpenFlow

Comparing the OpenFlow simulation with TCP/IP suite allows us to draw conclusions on performance evaluation. It is assessed by varying the same amount of additional traffic and link delay in both OpenFlow and TCP/IP simulation channels. In this experiment, we measure the mean values of measured RTT between the time-triggered nodes, both for the standard TCP and the OpenFlow.

The results in Figure 3 shows the mean RTT from measured scalar values for TCP and OpenFlow among different domains, and it is obvious that the performance of TCP clients in every domain has lower RTT values than OpenFlow when using the ping application. Another outcome of Figure 3 is that OpenFlow performances are affected by the placement of the central controller. For example, the last two domains, Perth and Sydney, where the controller was placed closer to Sydney then the result for this domain is better than Perth with 0.0905 seconds lower rate of the measured mean RTT.

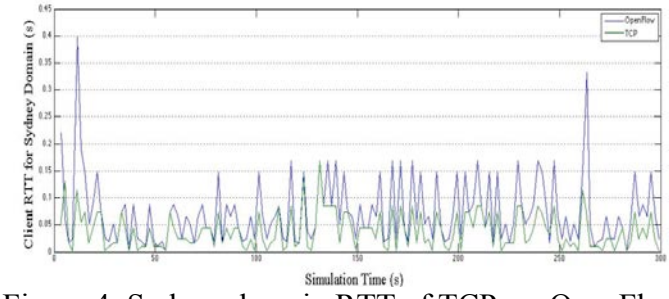



Figure 4: Sydney domain RTT of TCP vs. OpenFlow

Figure 4 and 5 show traces of the round trip times for both TCP and OpenFlow that was used to test the various algorithms. Moreover, they show the comparison in more details to prove the outcome of Sydney and Perth. Where OpenFlow Sydney domain RTT starts at 0.22 seconds, and the OpenFlow Perth domain RTT starts at 0.57 seconds. Thus, these differences are caused by the placement of the central controller. It is also evident that there are sudden spikes at the beginning and at the end of OpenFlow caused by connection establishment and termination. However, the performance of OpenFlow and TCP are slightly similar during 100 to 250 seconds of simulation run time. Because TCP takes less connection set-up time than OpenFlow, it overall performs with lower RTT values, which indicates how well the client can ping to others. This may confirm the assumption that TCP performs slightly better than OpenFlow with respect to mean round trip time (RTT), and performs faster with the same circuitry and not incurring major performance losses.

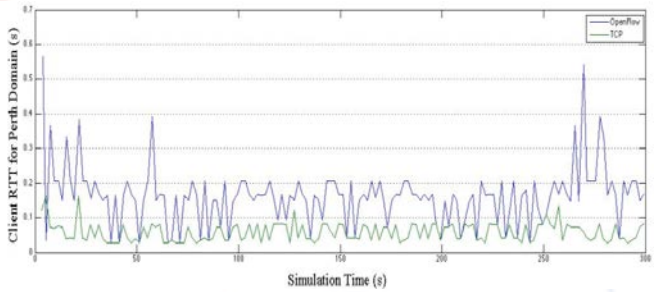



Figure 5: Sydney domain RTT of TCP vs. OpenFlow

**Proposed DAIM model and its implementation**

DAIM is a sustainable information model, which collects, maintains, updates and synchronizes all the related information. Moreover, the decision making ability within each device locally, on the basis of collected information, allows it to autonomically adapt according to the ever changing circumstances (Banjar et al., 2014b). The DAIM model structure is proposed with the hope that it addresses the limitation of previous network protocols such as Simple Network Management Protocol (SNMP), Common Information Model (CIM) and Policy-Based Network Management.

Proposed DAIM model will address the limitations of current approaches and future distributed network systems, creating an autonomic computing management strategy. The DAIM model approach will also satisfy the requirements of autonomic functionality for distributed network components like self-learning, self-adaptation and self-CHOP (configuration, healing, optimization and security). Each component can be adaptable according to any changed conditions of the dynamic environment without human intervention.

We are proposing that by creating a DAIM model on the networks we could give effect to what we are calling a Reactive Interpreter Network. So it would be a truly distributed computing environment, where these DAIM agents reside in the network elements, which would be OpenFlow switches (Pupatwibul et al., 2013). The actual values in the OpenFlow tables, reside in the OpenFlow switches, and would then be the properties of DAIM agents. These agents would then have to do the work of modifying or adapting their values so as to implement the requirements of the

network. So the whole DAIM model stretches across all these network elements and then could be thought of as reactive distributed interpreter that is interpreting the system requirements to enable the infrastructure to provide for the business needs.

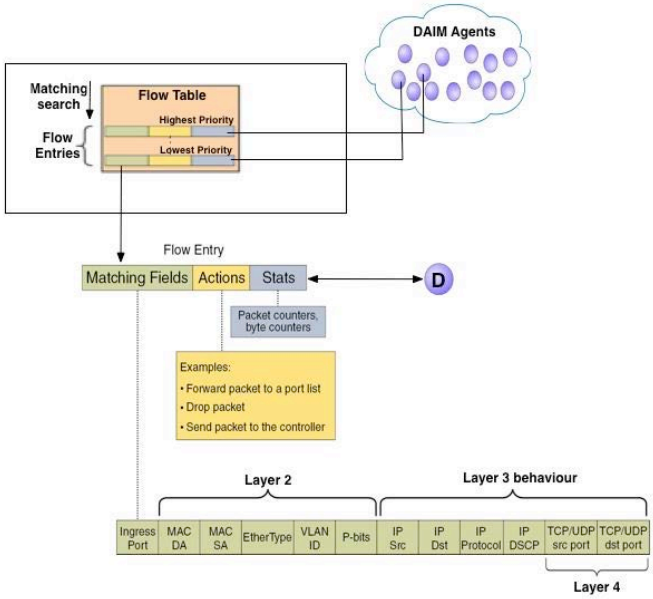



Figure 6: DAIM model integrated in OpenFlow switch

The DAIM cloud has a multi-agent operating system such as SPADE (Smart Python multi-Agent Development Environment) that can create, change, and terminate the intelligent DAIM agents. These agents have the responsibility to maintain their own values, and they can adapt and modify their own value according to the collected information. DAIM agents can make their own local decisions based on the system requirements (see Figure 6).

When the DAIM cloud receives an unmatched packet, it creates DAIM agents which can access and control network elements such as system requirement database, and other switches to determine the forwarding rules. The DAIM agents should be able to check this flow against system requirements and other policies to see whether it should be allowed, and if allowed the DAIM agent needs to compute a path for this flow, and install flow entries on every switch along the chosen path. Finally, the packet itself will be forwarded. The DAIM agents can provide a distributed environment where the network information is the property (values) of software agents residing in virtual machines that are distributed throughout the network elements. Therefore, the DAIM agents have the ingredients to implement autonomic behaviours.

## Conclusion

We have introduced TCP/IP and OpenFlow networks including the background, behaviour and architecture. We gave a thorough overview of the related works regarding simulation and emulation tools, which present advantages for researchers to use OMNeT++ over other tools. We also presented Round-Trip-Time as network metric and topologies of our study, followed by evaluating the performance of TCP/IP protocol suite in contrast with OpenFlow protocol. It has been evident according to the measurement outcomes that TCP/IP performs more effective than OpenFlow with lower RTT values and can send streaming UDP packets at a higher rate (Mbps).

Lastly, the proposed DAIM and its implementation have been introduced with the hope to resolve the scalability issues and develop autonomic behaviours in OpenFlow.

As for future studies, we aim to implement the DAIM cloud and extend OpenFlow structure based on intelligent agents to exchange information and install forwarding flow tables, which can be used in other distributed computing environment. Therefore, it could be applied to many different environments such as large data centers and road traffic systems.

**Acknowledgement**

This work is sponsored by the Centre for Real-Time Information Networks (CRIN) in the Faculty of Engineering & Information Technology at the University of Technology, Sydney (UTS).

## References

Banjar, A., Pupatwibul, P., Braun, R., Moulton, B. 2014a, "Analysing the performance of the OpenFlow standard for software-defined networking using the OMNeT++ network simulator," Computer Aided System Engineering (APCASE), 2014 Asia-Pacific Conference, pp.31-37.

Banjar, A., Pupatwibul, P. and Braun, R. 2014b "DAIM: a Mechanism to Distribute Control Functions within OpenFlow Switches." Journal of Networks9.01, pp. 1-9.

Fall K. and Floyd S., 1996 "Simulation-based comparisons of Tahoe, Reno and SACK TCP," ACM SIGCOMM Computer Communication Review, vol. 26, pp. 5-21.

Forouzan B. A., 2002 “TCP/IP protocol suite: McGraw-Hill”, Inc.

Gude N., Koponen T., Pettit J., Pfaff B., Casado M., McKeown N., and Shenker S., 2008 “NOX: towards an operating system for networks”. SIGCOMM Comput. Commun. Rev., 38(3):105–110.

Jeroen I., Heijenk G. and de Boer P. T., 2004 "TCP/IP modelling in OMNeT+.", Univers ity of Twente, The Netherlands.

Klein D. and Jarschel M., 2013 "An OpenFlow Extension for the OMNeT++ INET Framework," OMNeT++ 2013, Cannes France.

Lantz B., et al., 2010 "A network in a laptop: rapid prototyping for software-defined networks," in Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks, p.19.

McKeown N., Anderson T., Balakrishnan H., Parulkar G., Peterson L. Rexford, J., Shenker S. and Turner J., 2008 “OpenFlow: Enabling Innovation in Campus Networks” SIGCOMM Computer Communication Rev., Vol. 38, pp. 69-74.

NS-3 version 3.16. OpenFlow switch support. http://www.nsnam.org/docs/release/3.16/models/html/openflow-switch.html, Dec 2012.

ONF White Paper. Software-Defined Networking: The New Norm for Networks. April 2012.

ONF, 2012, “Open network foundation, openflow switch specification version 1.3.0 ( wire protocol 0x04 ) viewed 20-08-2012 www.opennetworking.org/images/stories/downloads/specification/openflow-spec-v1.3.0.pdf.

OpenFlow Switch Consortium. "OpenFlow Switch Specification Version 1.0.0." (2009).

Pupatwibul, P., Banjar, A. and Braun, R. 2013 "Using DAIM as a reactive interpreter for openflow networks to enable autonomic functionality." Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM. ACM, pp. 523-524.

Stewart R. and Metz C., 2001 "SCTP: new transport protocol for TCP/IP," Internet Computing, IEEE, vol. 5, pp. 64-69.

Sünnen D., 2011 “Performance Evaluation of OpenFlow Switches,” Semester Thesis at the Department of Information Technology and Electrical Engineering.

Varga A. and Hornig R., 2008 "An overview of the OMNeT++ simulation environment," in Proceedings of the 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops, p. 60.

Varga A., 2001 "The OMNeT++ discrete event simulation system," in Proceedings of the European Simulation Multiconference (ESM’2001), p. 185.

Varga A., 2010, OMNeT++ user manual. “OMNeT++ Discrete Event Simulation System”. Available at: http://www. omnetpp. org/doc/manual/usman.html.

Varga A., 2012 “INET Framework for the OMNeT++ Discrete Event Simulator”. http://github.com/inet-framework/inet.

Wang, S.-Y., Chou C.-L., and Yang Chun-Ming, 2013 “EstiNet openflow network simulator and emulator”, in Communications Magazine, IEEE. p. 110-117.

**Contact email:** 10926297@uts.edu.au, 11311103@uts.edu.au

iafor

# *Business Intelligence and Analytics to Prediction of Going Concern using Neuro-Fuzzy Approach*

Yee Ming Chen, Yuan Ze University, Taiwan, Republic of China(R.O.C)
Yu-Pu Chiu, Yuan Ze University, Taiwan, Republic of China (R.O.C)

The Asian Conference on Technology, Information & Society
Official Conference Proceedings 2014

**Abstract**
Rapid advances in technology, enterprise environmental changes and increasing competition has affected the risk of investment. Going concern prediction (GCP) is an important element in investor's evaluation. The evaluation of a enterprise 's going concern status is not an easy task. To assist auditors, going concern prediction models based on statistical methods such as multiple discriminant analysis and multiple linear regression analysis have been explored with some success. Nowadays, the business intelligence and analytics (BI&A) has attracted more and more attentions, which is required to manage immense amounts of data quickly. However, current researches mainly focus on the amount of data. In this paper, the other two properties of BI&A, which include the high dimensionality of data, and the dynamical change of data, are discussed. This study attempts to look at a different and more recent approach—Adaptive Neuro Fuzzy Inference System (ANFIS). ANFIS has effectively solved many large-scale, and dynamical problems. This study explores and compares the usefulness of logistic regression and proposed ANFIS in predicting an enterprise's going concern status. The sample data comprise financial ratios for 165 going concerns and 165 matched non-going concerns. The classification results from view of significance test and predictive accuracy which indicate the potential usefulness of BI&A in a going concern prediction context. These results also indicate that ANFIS shows acceptable performance in terms of accuracy and comprehensibility, and it is an appropriate choice for auditors to assess potential clients and as a means to identify non-going concern enterprises that might require further consideration.

**Keywords:** Business intelligence and analytics, Big data analytics, Neuro fuzzy approach

## Introduction

Business intelligence and analytics (BI&A) and the related field of big data analytics have become increasingly important in both the academic and the business communities over the past few decades. By carefully analyzing the application and data characteristics, researchers and practitioners can then adopt or develop the appropriate analytical techniques to derive the intended impact. Emerging analytics research opportunities can be classified into five critical technical areas—(big) data analytics, text analytics, web analytics, network analytics, and mobile analytics—all of which can contribute to to BI&A. Data analytics refers to the BI&A technologies that are grounded mostly in statistical analysis and data mining. Grounded in statistical theories and models, multivariate statistical analysis covers analytical techniques such as regression, factor analysis, clustering, and discriminant analysis that have been used successfully in various business applications. Since the late 1990s, various data mining algorithms have been developed by researchers from the business intelligence, algorithm, and database communities. Data mining in BI&A area aims to identify valid, novel, potentially useful and understandable correlations and patterns in data (Chung and Gray, 1999). More importantly, data mining techniques such as neural networks and soft computing provide a different approach to predictive modeling. Due to their prediction and classification capabilities, data mining techniques have been employed to facilitate the auditing process, to predict corporate performance, and to facilitate credit risk estimation. The going concern prediction (GCP) has become an important research in finance areas. In general, the objective of GCP is to develop models that can extract novel knowledge from previous observations and appraise corporate status. In this study, we have compared logistic regression and neuro-fuzzy approach in the context of going concern prediction. Both techniques are considered data mining techniques, but only logistic regression can be considered a traditional statistical method. Neuro-fuzzy approaches are developed for pattern recognition and prediction purposes in the area of business intelligence. Neuro-fuzzy approach via utilizing a large number financial data can be extracting, valuable and unknown knowledge dynamically. In this paper, the neuro-fuzzy approach of adaptive neuro-fuzzy inference system (ANFIS) model for prediction have been conducted in GCP area and the findings indicate that this technique is able to predict the going concern status of firms and accounting data are useful in GCP. This study explores ANFIS model and compares the usefulness of logistic regression (a traditional statistical method included in most data mining software as a data mining technique) in predicting a firm's going concern status. The sample data comprise financial ratios for 165 going concerns and 165 matched non-going concerns as a benchmark in auditor judgment. Results from this study will help a manager to keep track of company's performance and to identify significant problems and take efficient measure to reduce the coincidence of failure. The remainder of the paper is organized into the following three sections. Section 2 reviews the going concern literature and the third section ANFIS explores the potential usefulness of and discusses the comparative analysis of logistic regression in the going concern prediction context. Finally, the concluding section highlights the limitations of the study and suggests directions for future research.

## Background

The going concern prediction is a significant argument that a company will continue to operate for an indefinite period of time i.e. long enough to meet its objectives and fulfill its commitments. The prediction models of going-concern were discussed extensively in prior studies. Most of previous researchers used financial indicators to predict enterprise financial distress and bankruptcy. The literature on going concern prediction dates back to 1976, shortly after the issuance of Statement of Auditing Standards (SAS) No. 2, which was the first SAS to detail specific considerations for the auditor's assessment of a company's going concern status. The first study about going concern prediction published shortly after the issuance of SAS No.2 in 1974 by McKee (1976). Subsequent research applied statistical methods analysis to test going concern predicting models (e.g. Altman, 1968; Collins & Green, 1972; Balcaen,et al., 2006). Common tools for statistical analysis include linear regression and logistic regression. But, sstatistical models have certain distributional hypothesis that financial statement data do not always fit. Hence some non-parametric techniques have been developed to overcome the constraints of traditional statistical models. Most of them belong to BI&A domain such as decision trees, artificial neural networks (ANN), and support vector machines (SVM). Most researchers use one of the techniques to compare the prediction performance with other techniques for a specific data set (Koh et al., 1999; Shin et al., 2005; Zheng et al., 2007). There are a number of arguments which promote the consideration of the hybrid analysis model for going concern prediction using some BI&A techniques. In addition, these studies have shown that BI&A models outperform traditional statistical models. Kumar and Ravi (2007) provide a detailed review of these models in the domain of going concern prediction.

The going concern prediction studies often use samples of bankrupt firms versus non-bankrupt firms to assess model accuracy. While the ability of firms to continue in business is a concern for firms in any country, most studies have developed going concern prediction models for U.S. firms. In this research, the sample data for constructing the going concern prediction models is select from Predicast's F & S Index of Corporate Changes for the years 1980 to 1992. It comprises 165 going concerns and 165 matched non-going concerns in the US based on a compilation of bankruptcy data from numerous newspapers and periodicals. Studies usually report prediction accuracies separately for non-going concern firms (those which are not likely to survive) and going concern firms. In addition, the going concern prediction literature refers to Type I and Type II errors, which are applicable to going concern prediction models. Type I errors are the misclassification of non-going concern (i.e. bankrupt) firms as going concerns. Type II errors are the reverse – going concern firms misclassified as non-going concern firms. Type I errors are generally considered more costly than Type II errors for several reasons including loss of business (audit clients), damage to a firm's reputation, and potential lawsuits/court costs (see for example Koh [1987]). Therefore, the predictive accuracies discussed here refer to the accuracies obtained for non-going concern firms, unless the results were not presented separately for non-going concern and going concern firms.

## Construction of going concern prediction models

In this paper, we have investigated BI&A techniques. In this section, we used the logistic regression and Adaptive Neuro Fuzzy Inference System (ANFIS) models to

predict the going concern and non-going concern firms. The sample data for constructing the two going concern prediction models is taken from Predicast's F & S Index of Corporate Changes for the years 1980 to 1992. Based on the sample data, the objective is to construct a prediction model for the going concern status of firms (a dichotomous dependent variable) based on the six financial ratios (independent variables).The following independent variables were applied to the going-concern companies to select the sample that fits the current study:

(1) Quick assets to current liabilities (QACL).
(2) Market value of equity to total assets (MVTA).
(3) Total liabilities to total assets (TLTA).
(4) Interest payments to earnings before interest and tax (IEBT).
(5) Net income to total assets (NITA).
(6) Retained earnings to total assets (RETA).

To derive the going concern prediction model, the sample data is partitioned into the following two data sets: a construction or training sample (comprising approximately 70 percent of the original samples) and a validation or testing sample (comprising the remaining 30 percent of the original samples).

**Logistic regression**

This study used logistic regression to build the prediction model of going-concern, logistic regression is an appropriate traditional statistical method to use. The logistic regression results are summarized in Table 1. As can be seen, the logistic regression model indicate a good fit. At a 0.05 level of significance, QACL ( p-value= 0.0114), TLTA ( p-value = 0.0001) and RETA ( p-value= 0.0004) are statistically significant. The numerical sign of the respective coefficient shows that a higher level of liquidity (QACL), a lower of leverage (TLTA) and a higher level of profitability (RETA) are associated with a higher probability of continued going concern status. Based on the results, it can be concluded that the logistic regression going concern prediction model is acceptable.

Table 1. Logistic regression for prediction model of going concern

| Variable | Coefficient | P-value |
|---|---|---|
| **Intercept** | 13.64 | 0.0002 |
| **QACL** | 2.74 | 0.0114 |
| **MVTA** | 0.85 | 0.2342 |
| **TLTA** | -23.41 | 0.0001 |
| **IEBT** | -0.45 | 0.4215 |
| **NITA** | 3.74 | 0.0734 |
| **RETA** | 17.89 | 0.0004 |

**ANFIS Model**

The adaptive network-based fuzzy inference systems (ANFIS) is used to solve problems related to parameter identification. This parameter identification is done through a hybrid learning rule combining the back-propagation gradient descent and a

least-squares method. ANFIS is basically a *graphical* network representation of Sugeno-type fuzzy systems endowed with the neural learning capabilities. The network is comprised of nodes with specific functions collected in layers. ANFIS is able to construct a network realization of IF / THEN rules. Consider a Sugeno type of fuzzy system having the rule base

(I). If $x$ is $A_1$ and $y$ is $B_1$, then $f_1 = c_{11}x+c_{12}y+c_{10}$
(II). If $x$ is $A_2$ and $y$ is $B_2$, then $f_2 = c_{21}x+c_{22}y+c_{20}$

Let the membership functions of fuzzy sets *Ai, Bi, i*=1,2, be , $\mu_{Ai}$ $\mu_{Bi}$.

In evaluating the rules, choose *product* for T-norm (logical *and*).

(1). Evaluating the rule premises results in

$$w_i = \mu_{A_i}(x)\mu_{B_i}(y), \quad i = 1,2 \tag{1}$$

(2) Evaluating the implication and the rule consequences gives

$$f(x,y) = \frac{\sum_i w_i f_i}{\sum_i w_i} \tag{2}$$

This can be separated to phases by first defining

$$\overline{w}_i = \frac{w_i}{w_1 + w_2} \tag{3}$$

(3) Then $f$ can be written as

$$f = \overline{w}_1 f_1 + \overline{w}_2 f_2 \tag{4}$$

All computations can be presented in a diagram form. ANFIS normally has 5 layers of neurons of which neurons in the same layer are of the same function family.

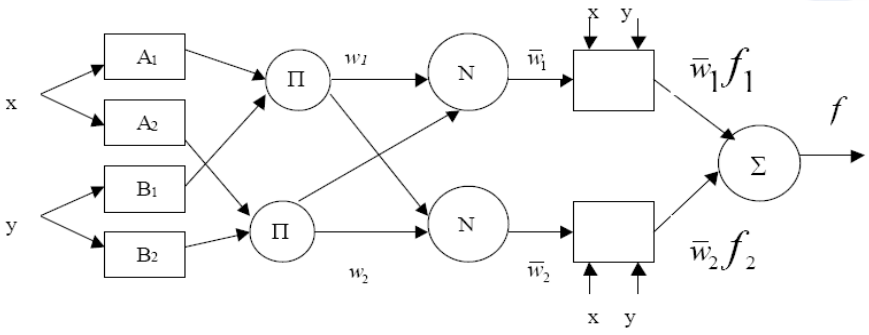



Figure 1 Structure of the ANFIS network.

We get our going concern prediction model on the Predicast's F & S Index of Corporate Changes database. The proposed method for Predicast's F & S Index of Corporate Changes ANFIS model is depicted in Fig. 2 which shows actual and predicted dichotomous value (going concern or otherwise) using MathWorks MATLAB as a software support. The ANFIS model in this study was simulated

using MATLAB (R2012b). ANFIS with two input membership functions, generalized bell membership function and our linear generating function, were trained.

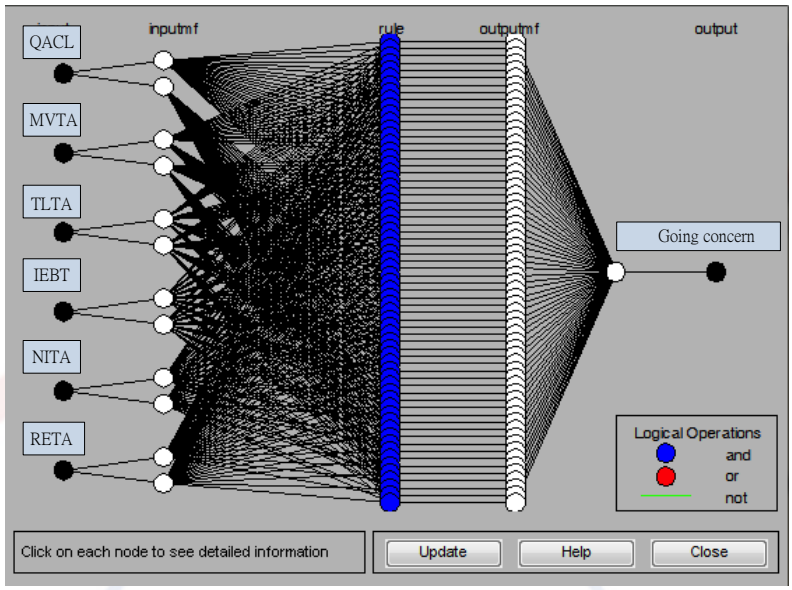



Figure 2 Going concern prediction ANFIS model implemented using Matlab software

The data flow and train process of ANFIS model is shown in Fig. 3.

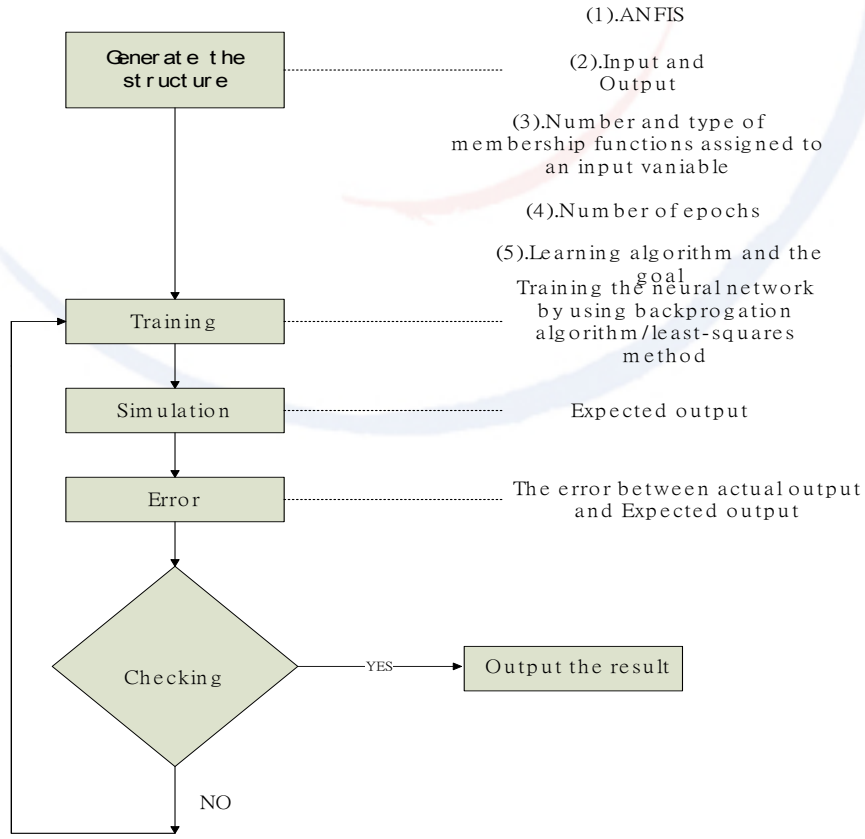



Figure 3 Data flow and train processing in ANFIS model

The data flow and train process for ANFIS model can be divided into two stages as shown in Fig. 3. In the first stage, the generated of the ANFIS structure (depicted in Fig. 2). In the second stage, ANFIS training process starts by obtaining a data set (input-output data pairs from the Predicast's F & S Index of Corporate Changes database) and dividing it into training and checking data sets. Training data constitutes a set of input and output vectors. The data is normalized in order to make it suitable for the training process. The training data set is used to find the initial premise parameters for the membership functions using backpropagation algorithm. The consequent parameters are found using the least-squares method. The train process is terminated when the error becomes less than the threshold value. Then the checking data set is used to compare the model with actual data. Figure 4, for instance, shows the error curves for 100 epochs of ANFIS training. The green curve gives the training errors and the red curve gives the checking errors. The minimal checking error occurs at about epoch 45, which is indicated by a circle. Notice that the checking error curve goes up after 50 epochs, indicating that further training overfits the data and produces worse prediction.

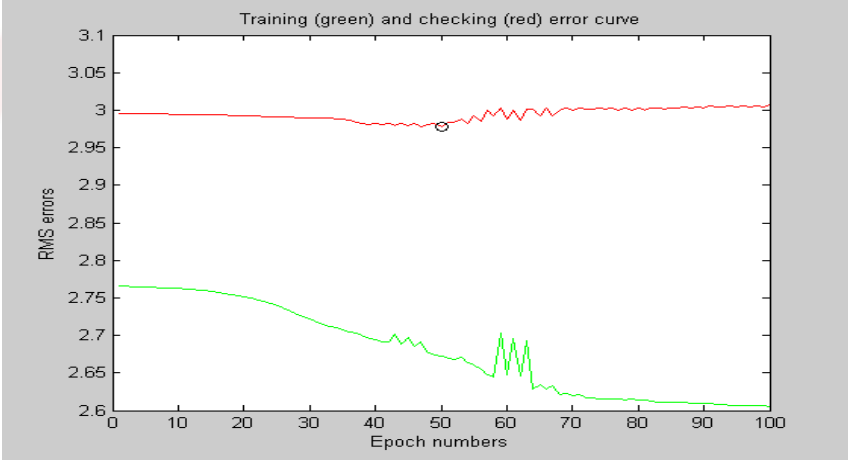



Figure 4 ANFIS training and checking errors

After training association rules in the form of IF-then rules are generated and extracted. an example of initial and final (after rule extraction) decision surfaces are given in Fig. 5. The input-output surface of the ANFIS model is shown in the plot below.

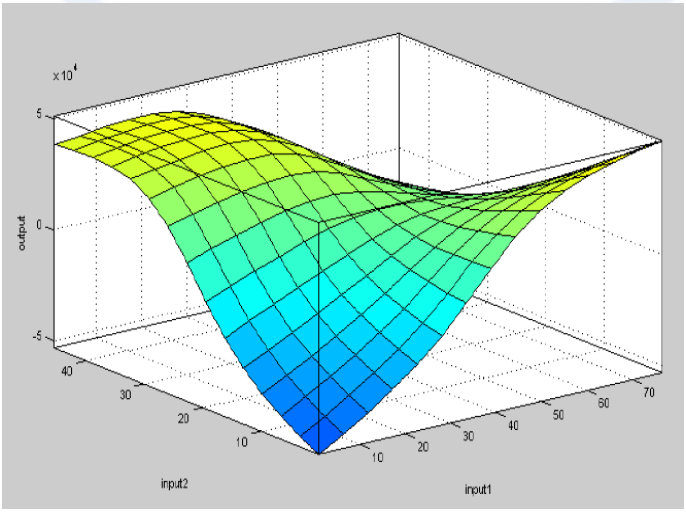



Figure 5 Final decision surface for input 1 and input 2

The predictive capability of using neural network and ANFIS approaches are compared using statistics in the next subsection.

**Predictive Results**

The two models of logistic regression and ANFIS that were constructed based on BI&A. We use the Predicast's F & S Index of Corporate Changes test to examine whether the predictive performance of the ANFIS method is significantly higher than of traditional statistical method. The predictive ability of a model can also be impacted by whether the results are from tests of a hold-out sample. Hence, to provide an unbiased assessment of the performance of the two prediction models, the validation sample is used 10-fold cross-validation. This method splits the data into two subdivisions: a training set and test set. Quality of the prediction assessed on the test set. In 10-fold cross-validation the data is firstly partitioned into 10. Then, 10 iterations of training and test are done such that in each iteration a different fold of the data is held-out for validating while the rest 9 folds are used for learning and 10 outputs from the folds can be averaged and can produce a single prediction. The validation predictive results of each models listed in Table 2.

Table 2 Predeiictive accuracy(%) of hold-out data

| Data sets | ANFIS | | | Logistic Regression | | |
|---|---|---|---|---|---|---|
| | Accuracy | Error Type I | Error Type II | Accuracy | Error Type I | Error Type II |
| 1 | 100.0 | 0 | 0 | 86.51 | 0 | 28.46 |
| 2 | 100.0 | 0 | 0 | 93.15 | 0 | 16.51 |
| 3 | 93.35 | 0 | 12.52 | 80.05 | 0 | 38.42 |
| 4 | 100.0 | 0 | 0 | 86.55 | 12.51 | 14.16 |
| 5 | 100.0 | 0 | 0 | 86.13 | 0 | 26.23 |
| 6 | 100.0 | 0 | 0 | 100.0 | 0 | 0 |
| 7 | 100.0 | 0 | 0 | 100.0 | 0 | 0 |
| 8 | 98.14 | 0 | 0 | 92.57 | 0 | 14.32 |
| 9 | 98.14 | 3.14 | 9.36 | 93.15 | 0 | 12.53 |
| 10 | 100.0 | 0 | 0 | 80.11 | 0 | 38.1 |
| Min | 93.35 | 0 | 0 | 80.05 | 0 | 0 |
| Max | 100.0 | 3.14 | 9.36 | 100.0 | 12.51 | 38.42 |
| Mean | 99.15 | 0.31 | 2.19 | 89.82 | 1.25 | 18.87 |
| Median | 100.0 | 0 | 0 | 89.56 | 0 | 15.41 |
| Variance | 2.018 | 0.9420 | 4.432 | 6.79 | 3.75 | 13.03 |

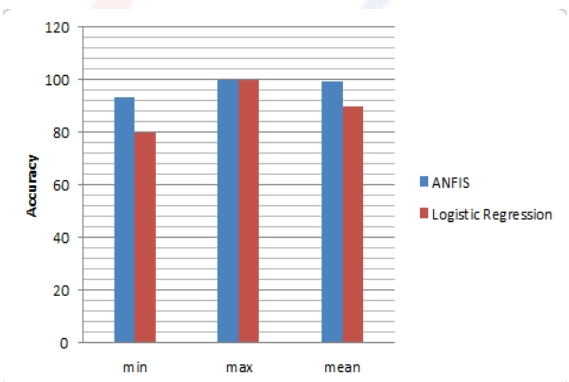
The validation results also indicate that the ANFIS model (accuracy =99.15%) outperforms the logistic regression model (accuracy=89.82%). Hence, if only one model is to be selected, it will be the ANFIS model as it is expected to perform better on new observations as compared to the traditional models.

Type I error is the probability that a company with non-going concern status to be predicted as a company with going concern status and Type II error is the probability that a company with going concern status to be predicted as a company with non-going concern status. In holdout data type I and II error are also equal to 0.31 and 2.19 percent in ANFIS model and 1.25 and 18.87 percent for obtained model by logistic regression. Determine which of the models that are more applicable than others in GCP, we make the significance test between two models. As shown in Table 5, the result of the Predicast's F & S Index of Corporate Changes test at 5% level indicates that there are significant differences between the two models in GCP.

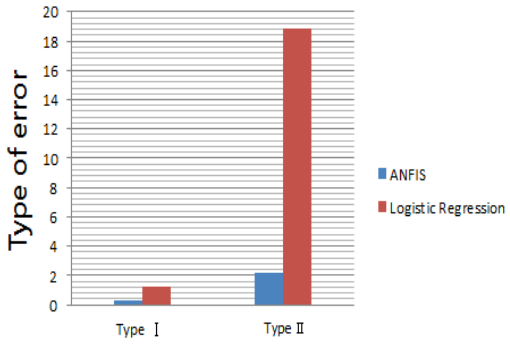
Table 5 Results of significance test between two models

| Models | Logistic Regression |
|---|---|
| **ANFIS Model** | $-2.569^{a}(0.011)^{b}$ |

**[a]= *t* statistic;[b]= p-value**




(a) The map of mean accuracy of both models




(b) The map of mean of difference types of errors

Figure 6. Going concern prediction experiment al results

From Fig. 6(a), we can be understood that ANFIS outperforms logistic regression on the mean accuracy. Fig. 6(b) shown the related to these two types of errors are very different. The resulting shown incorrectly predicted a company with going concern as a company with non-going concern status (Type II error) is much larger than the Type I error (incorrectly predict a company with non-going concern as a company with going concern status). Type I errors are generally considered more costly than Type II errors for several reasons including loss of business.

## Conclusions

In recent years, BI&A data mining has gained widespread attention and increasing popularity in the business world. Successful neuro-fuzzy approach applications have been reported and recent surveys have found that ANFIS has grown in usage and effectiveness. The main purpose of the paper is to predict company's going concern by using the ANFIS model and to show its adequacy experimentally. The results of the experiments show that the suggested model predicts company's situation more

accurately in comparison to logistic regression prediction model, especially evaluating bankrupt companies. It can be stressed that a neuro-fuzzy approach in going concern prediction is a potentially powerful alternative or complement to the more commonly used statistical methods. The development of going concern status analysis models to predict business failures can be thought of as early warning systems, which proved to be very helpful for managers, and relevant authorities who can prevent the occurrence of failures.

## References

Altman, E. I. (1968). Financial ratios, discriminant analysis and prediction of corporate bankruptcy. The Journal of Finance, 23, 589 – 609.

Aziz, M., Dar, H. (2006). Predicting corporate bankruptcy- where we stand? Corporate governance journal, 6(1), 18-33.

Balcaen, S., & Ooghe, H. (2006). 35 years of studies on business failure: An overview of the classic statistical methodologies and their related problems. The British Accounting Review, 38(1), 63 – 93.

Collins, R. A., & Green, R. D. (1972). Statistical methods for bankruptcy forecasting. Journal of Economics and Business, 32, 349 – 354.

Koh, H. C. and Tan, S. S, (1999). A neural network approach to the prediction of going concern status. Accounting and Business Research, 29 (3), 211-216.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques – A review. European Journal of Operational Research, 180(1), 1 – 28.

McKee, T.(1976). Discriminant prediction of going concern status: A model for auditors. Selected Papers of the AAA Annual Meeting .

Shin, K. S., lee, T. S., Kim, H. J, (2005). An application of support vector machines in bankruptcy prediction model. Expert system Application, 28(1), 127-135.

Zheng, Q., J. Yan-Hui.(2007), Financial Distress Prediction on Decision Tree Models , IEEE, Service Operations and Logistics, Informatics, International Conference.

Corresponding author: chenyeeming@saturn.yzu.edu.tw

iafor

# *Use of Social Media in Academic Environment: a New Platform for Enhancing Independent and Collaborative Learning*

Behdja Boumarafi, University of Constantine 2, Algeria

The Asian Conference on Technology, Information & Society
Official Conference Proceedings 2014

**Abstract**
In past two decades social media technologies generated deep transformation in societies by capturing worldwide audiences. Social network sites (SNS) that were initially used for socializing, and entertainment have in recent years experienced a rapid increase in participation that has been accompanied by a diversification of purposes and usage patterns. They expanded their reaches into learning activities providing an open space to create content, share ideas and engage with universal knowledge. They are offering a new perspective to students' engagement in collaborative learning to expand their own learning outside the classroom boundaries. As education institutions are embracing social media technologies there is a need to optimize the positive effect of such technologies to bring them into pedagogy to make instruction and learning active and applicable to the cyber environment of the new millennium. Social networking appears very helpful in building academic groups to achieve independent and collaborative learning. Students learn to communicate and collaborate electronically, conduct research online, create digital objects in form of text documents, images, email messages, online discussion and create their e-portfolio. This requires rethinking the way pedagogical activities are carried out in the networked environment and assessing their effect on instructors and learners to create the best learning model in line with this new educational paradigm.

The main purpose of this paper is to investigate Algerian students' engagement in academic interpersonal relationships. It examines students' interpretation of collaborative learning to understand the developmental implications of academic achievement. The author seeks to advance the use of instructional interactions in two ways: i: enhancing the academic validity of social networking, ii: making the link to students' collaborative learning explicit. An online questionnaire is used to collect data from 300 students and statistical analysis of participants' responses show that social media are used for a multitude of purposes including communication and entertainment. They also use SNS to have a presence in the virtual world and improve their personal image. Facebook topped the use followed by YouTube, Twitter, LinkedIn, and the least used network is MySpace.

**Keywords**: social media, social networking, collaborative education, active learning, independent learning, academic achievements, Algerian students.

iafor
The International Academic Forum
www.iafor.org

## Introduction

The unprecedented developments of information and communication technology (ICT) and related digital technologies have ushered in many changes in all spheres of life. They have changed the way people communicate, work and study. Furthermore, the speed at which such technologies are penetrating education institutions and offering new ways and tools for the delivery of knowledge across the globe making the cyber-space a borderless learning sphere. By the adoption of social media in academia, instructors are shifting some of their teaching activities to students seeking their involvement in enhancing self and collaborative learning. Documented literature that attempts to shed light on social media adoption in learning indicates an increase in the use and popularity of social network sites among students. Facebook, twitter, LinkedIn, Skype, YouTube, MySpace, Google+… are making a considerable impact as essential tools to promote exchange of knowledge and to create virtual academic communities. The internet revolution is creating a new culture in learning and instruction activities. It is very clear that education industry worldwide is in the midst of a revolution caused by the evolving technologies such as the web 2.0 and the advent of web 3.0 allowing students to create content, exchange ideas and share knowledge. So much so that in the last few years there has been extensive discussion and heated debate exploring social media in journal articles and conferences and much of this discussion has focused on developing a clearer understanding of the capabilities of such technology as a new platform for enhancing students' independent learning and how much academic achievement social networking has yielded and whether it could be used as a new pedagogical tools outside the classroom.

As education institutions are embracing social networks there is a need to optimize the positive effect of such technologies to bring them into pedagogy to make instruction and learning active and applicable to the cyber environment of the new millennium. In recent years, more and more education institutions are making a presence in social networks such as blogs, twitter, YouTube, Facebook among others to create effective ways for the establishment of collaborative and interactive online learning system. Therefore, technology-driven culture is featuring prominently in all pedagogical activities. Hence, the focus of learning and instruction needs to be viewed from a newer perspective "without gathering students and teachers in the same physical space" (Maney, 2009). By and large, learning is no longer tied to a particular location and a particular time table. For a generation immersed in a world of evolving technologies where internet applications, specifically the web 2.0 and 3.0 is having a considerable impact on the way people work, study, play, and communicate. Social network sites promise the opportunity to motivate students to constantly quest for new learning experience out of the physical locations of the university encouraging them to create and share ideas rather than simply memorizing lecture notes. The use of social networks in education has created a paradigm shift from emphasis on teaching content to helping students develop the ability to create content tailored to their learning needs and chare it with peers. Basically, social network sites provide open avenues for "collaborative education which in essence creates a synergy by bringing together technologically learners and instructors with a shift towards less instructor led teaching to a greater student's involvement" (Boumarafi, 2010). It is therefore, pertinent to say that the use of social network sites extend far beyond their traditional purpose of communication and entertainment to promote students' self-reliance in learning through enquiry and sharing. This

underscores the need to empower students to take charge of their own learning taking place anytime anywhere. This requires a rethinking of the way core pedagogical activities are carried out and assess their effect on both learners and instructors. The purpose of this study is to provide a better understanding of how students at the University of Conantine 2 (Algeria) are investing their technological skills and learning time, in using social networks for better academic achievements and to examine factors affecting their use. The study will build upon the existing body of documented literature that seeks to explore the use of social media in the academic environment. To explore such understanding further, the following objectives were set forth:

- Seek Algerian students' insights about the use of social network sites in learning;
- Identify which social network sites are more popular among students;
- Identify the raisons for social networking.
- Identify which personal information users reveal when using social media.

## Literature Review

Boyd and Ellison (2008) traced the history of social network sites (SNs) use back to 1997 when sixdegrees.com, was launched. Since then several social networks were established between 1998 and 2000. However, it was not until the introduction of Friendster, Myspace and Facebook that SNs started to grow in popularity and gain significant mainstream and worldwide popularity (comScore, 2007). The use of social media in academic environment was explored by Jabr (2011) who argued that people use social media for interacting to exchange information, brainstorm ideas, search for new friends, and share everyday news. Back in 1996, Blake and Sikkal declared that in the academic environment, new technological developments specifically Web 2.0 applications will group students into virtual communities whose homogeneity in interest and intellect will create a synergy to learn and benefit from their heterogeneous cultural background and experience. In a similar context, Ellison, Steinfield & Lampe, (2007) observed that Facebook supports resource sharing by establishing the social foundation between students and their peers. In essence the advent of social networking technology is also the advent of new learning systems. There is a whole host of surveys that point to the increased use social network sites by academic community and their impact on instruction and learning activities. Just as social networks extended the function of communication, they also extended their usefulness to learning and instruction allowing for better "communication, collaboration, community creativity, and convergence". (Friedman and Friedman, 2008). The best example of this is the increasing trend towards what is today called collaborative and interactive learning where students' self-governed and problem-solving activities are considered the focal point of a new learning process. For example, Tiryakioglu and Erzurum (2011) moved the focus to facebook as an educational tool. They reported that the majority of participants they surveyed have a Facebook account. Most of them used it as a communication tool. 80% agreed that it contributed to the communication between classmates. Results of their study also indicate that most faculty members are willing to use Facebook to create academic groups to share information about their courses. Al Muitairi's (2013) study also highlighted that Facebook is used as a supporting tool offering faculty at the University of Kuwait a substitute channel to enrich their educational experience through a communication channel for both students and educators. Jabr (2011) used

online questionnaire to explore the use of social networks by students at Sultan Qaboos University in Oman. The study revealed that 70% of the 650 respondents use Facebook, 26% have a MySpace account, 9% use Twitter and only 5% use LinkedIn. The study confirmed that these are mostly used for communication and socializing purpose at a percentage of 91 and only 23% use social networking for assignment preparation.

Some researchers are not convinced that social networking can be used as a medium of education. For example, results of a survey conducted by Malesky and Peters (2012) show that 40% of students and 30% of faculty believe that it is inappropriate for professors to use SNs. Other surveys have painted a different picture. For instance Reuben (2008) contends that there is a great potential in education for Facebook and YouTube. SNs have been utilized by institutions of higher education to connect with, and disseminate information to current students, potential students, and alumni. Universities have links to Facebook and Twitter on their homepage; they use them to relay information to students (Malesky and Peters, 2012). In retrospect it is important to learn how to integrate evolving technology into learning strategies; not just for technology's sake, but for the added value that these tools already familiar to learners provide Brotherton (2011). Social media can cause negative feelings and reduce members' life satisfaction. A survey of 600 facebook users reported that more than one third developed feelings of frustration and envious when reading friends' postings, checking news feeds and browsing photos related to friends' travel and leisure posted on the platform. This may cause a negative effect on users' satisfaction with their own life. (Krasnova and Buxmann, 2012)

Despite the contradictory survey results, what is clear is that a considerable portion of the digital era students consider the traditional mode of learning and instruction no longer the most appropriate model of education. However, the growing use of social network sites by students for learning has increased concerned about the quality of knowledge exchanged among them and how do they assess its credibility. Although social network sites tools are not created for educational purposes, Dalsgaard (2005) argues that they can be used to support learning. Using social software can help facilitate an approach to e-learning which differs from using learning management systems and which better supports self-governed, problem-based and collaborative activities. Basically, the purpose is to provide students with tools which they can use to solve problems on their own and/or in collaboration with other students. Battouche (2012) conducted a seminal study to investigate the use of social networks by Algerian youth aged 8 to 24. He found that 84% of the surveyed population used Facebook, 8% used twitter and only 4% used MySpace. 37% of respondents indicated that making new friends was the main reason for using social networks, 40% cited their usefulness in acquiring new knowledge, 11% mentioned effective communication and 18% networking. The results of the study also show that 65% of the surveyed population benefited from social networks in improving their foreign language skills, 25% in strengthening national identity and belongingness and 11% in acquiring the ability to convince in debates.

## Research Methods

### Data collection

To collect data for this study, a questionnaire instrument was designed based on the previous literature. The questionnaire used a five-point Likert scale ranging from (1) completely inappropriate as a learning tool to (5) completely appropriate. The statistical package for social sciences (SPSS) was used to analyze the data. Four students and two faculty members pre-tested the questionnaire. Based on their suggestions, the questionnaire was revised to its final form.

### Survey population

Participants of this study consisted of master students at the Institute of Library and Documentation Science at the University of Constantine 2. A request for participation along with 300 questionnaires was e-mailed to participants through the "Infocomdz" group. This is a discussion group created by students of the University of Constantine 2 and using it to communicate and share information and knowledge. Respondents were asked to complete the questionnaire which included questions on type of social network used, frequency of use of social network sites, purpose for the use of social network sites, perception about the usefulness of social network sites, and drawbacks of social network sites. 140 usable questionnaires were returned; a response rate of 46.66%.

## Results

### Participants

All respondents were master students at the Institute of Library and Documentation Science. Female student participants (n=90; 64.28%) out- numbered male participants (n=50; 35.71%). In actual fact the number of female students outnumbers the male students' intake in general in Algerian Universities. The age group of students is 21-28, those who grew up as active users of different internet applications i.e. the internet generation. All participants (n=140; 100%) reported using social network sites (SNs) for different purposes and some of them have more than one account.

### Type of social network preferred by students

Respondents were asked which social network they prefer to use. Their preferences are summarized in table 1. As expected, all respondents 140 (100%) gave the top rank to Facebook at a mean of 4.95. Previous studies also identified Facebook as the most frequently used SNs (Tiryakioglu and Erzurum, 2011; Jabr, 2011; Battouche, 2012; Al Muitairi, 2013). Twitter is used by 70 (50%) respondents as the second most popular social network site with a mean of 3.95. Respondents put YouTube in third position in terms of importance with 40 (28.57%) users and a mean of 3.78. Other social network sites scored lower means i.e: 2.75 for Skype that respondents use to communicate with family and friends abroad. 2.01 for LinkedIn. MySpace is ranked last with only 5 (3.58%) respondents at a mean of 1.40.

**Table 1: Preference of type of social network sites (n=140)**

| Type of social network | Mean | Standard Deviation | Number of users | percentage | Rank |
|---|---|---|---|---|---|
| Facebook | 4.95 | 2.98 | 140 | 100 | 1 |
| YouTube | 3.95 | 1.83 | 70 | 50 | 2 |
| Twitter | 3.78 | 2.78 | 40 | 28.57 | 3 |
| Skype | 2.75 | 2.04 | 15 | 10.70 | 4 |
| LinkedIn | 2.01 | 1.97 | 10 | 7.15 | 5 |
| MySpace | 1.40 | 2.52 | 5 | 3.58 | 6 |

**Purpose for using Social network sites**

Respondents were asked to indicate the purpose for using social network sites. According to the results in Table 2, it is clear that social network sites is primarily used as a communication tool for exchanging e-mails, receiving the top rank with the highest mean 4.5. Keeping in touch with family and friends is ranked second at a mean of 4.1. Making new virtual friends scored 3.98, followed by sharing photos, tracing an old friend, with means of 3.95 and 3.85 respectively. These results match the findings of Gentzler and Oberhauser (2011) who reported that students use social networks mostly for communication and socializing. Respondents use social networks to discuss group project works, share assignments and course work as well as files and lecture notes also scoring high means ranging from 3.78 to 3.35 and creating content and exchanging ideas at means of 2.98 and 2.50 respectively. This indicates that the surveyed population of this study uses social networks not just for socializing and communication but also for academic activities. Respondents seem to be interested in becoming global citizen by having a presence in the virtual world. Surprisingly, the study found that joining discussion forums, searching for a job and downloading videos and audios were the least reasons for using social network sites among the surveyed population. The most likelihood is that participants are using other websites for online discussion and looking for jobs in other job announcement sources i.e. newspapers and job sites.

**Table 2: purpose for using social network sites (n=140)**

| **Purpose** | **Means** | **Standard deviation** | **Rank** |
|---|---|---|---|
| Communicate with others through e-mail | 4.5 | .62 | 1 |
| Keep in touch with family and friends | 4.1 | .98 | 2 |
| Make new virtual friends worldwide | 3.98 | .74 | 3 |
| Share photos | 3.95 | 1.00 | 4 |
| Trace old friends | 3.85 | 1.05 | 5 |
| Have a presence in the virtual world | 3.83 | .93 | 6 |
| Discuss group project works | 3.78 | 1.18 | 7 |
| Share assignments and course work | 3.50 | 1.15 | 8 |
| Share files and lecture notes | 3.35 | 1.08 | 9 |
| Create content | 2.92 | 1.99 | 10 |
| Exchange ideas | 2.50 | 1.87 | 11 |

| | | | |
|---|---|---|---|
| Join discussion forums | 2.35 | 2.50 | 12 |
| Searching for a job | 2.25 | 2.90 | 13 |
| Download videos and audios | 2.00 | 1.89 | 14 |
| Make a presence in the cyber-space | 1.98 | 2.00 | 15 |

**Students' perception about the usefulness of social network sites**

The study intends to explore the students' opinion about the usefulness of social network sites. Results in table 3 show that SNs are primarily used for building relationships with a mean of 4.90 followed by "improve communication skills with 4.50 and develop social skills ranked third at a mean of 4.30. These findings support the results in table 2 that indicate that social networking is still more popular among Algerian students for communication and socializing. Respondents perceive social networking appropriate for improving their foreign language skills. This could be the reason for "making new virtual friends worldwide" being the fourth most important reason for using social networks as shown in table 2. Participants seem a little less clear about the using SNs to ''share ideas, improve learning, make contact with faculty easier, and enhance academic achievement" giving them means less than the mid-point 3. This could be attributed to the fact that the use of social networks in academic activities is still new and not perceived as a very important tool for academic purposes (Al-Muitairi, 2013).

**Table 3: Usefulness of social network sites (n=140)**

| **Statement** | **Means** | **Standard deviation** | **Rank** |
|---|---|---|---|
| Build relationships with others | 4.90 | 1.96 | 1 |
| Improve communication skills | 4.50 | 1.65 | 2 |
| Develop social skills and modify one's behavior | 4.30 | 2.05 | 3 |
| Improve foreign language skills | 4.05 | 1.15 | 4 |
| Improve one's personal image | 3.98 | 2.01 | 5 |
| Self regulated learning | 3.90 | 2.00 | 6 |
| Share ideas and promote creativity | 3.85 | 1.99 | 7 |
| Create Peer to peer learning communities | 3.80 | 1.15 | 8 |
| Personalization of learning content | 3.75 | 1.75 | 9 |
| Improve learning in general | 3.50 | 1.70 | 10 |
| Make contact with faculty easier | 2.00 | 1.22 | 11 |
| Enhance academic achievement | 1.30 | 1.90 | 12 |

**Drawbacks of using social network sites**

Participants were asked to indicate the problems and drawbacks related to the use of social network sites. Their answers as reported in table 4 indicate that students are very much concerned about the violation of their privacy rating it first with a mean of 3.73. They report that social networking is time consuming at a mean of 3.58. Participants seem to believe that social media is inappropriate for formal academic

activities. This finding is supported by previous literature which indicated that social networks are used for some sort of informal academic purpose; for revision, arranging group or project work, often initiated by students themselves and not part of a formal requirement of a course (Downes, 2004). Undoubtedly, this raises concerns about quality control of information posted by students on social networks. Participants reported "discouraging face-to-face communication" ranking it before last with the lower mean score of 2.43. Finally, participants believe that having a presence in the virtual world may turn them into schizophrenic leading two separate lives; one virtually and one in the real world and it is easy to lie about one's life. Some respondents mentioned the use of YouTube and Skype where they can see whom they are networking with in a virtual environment.

**Table 4: Social network sites drawbacks (n=140)**.

| Statement | Means | Standard deviation | Rank |
|---|---|---|---|
| Violates privacy | 3.73 | 1.18 | 1 |
| Time consuming | 3.58 | 2.18 | 2 |
| Concerned about the quality of information | 3.35 | 2.03 | 3 |
| Not appropriate for formal academic activities | 2.87 | 1.94 | 4 |
| Discourages face-to-face communication | 2.43 | 1.82 | 5 |
| Can cause schizophrenia | 2.25 | 1.78 | 6 |
| Easy to lie about one's situation | 2.00 | 1.55 | 7 |

## Conclusion

Professional literature indicates an increase in the use of social network sites highlighting their popularity among students. The literature also shows that social networks are primarily used for communication and socializing. They are gradually making inroads as educational tools, and seem to have the potential to support collaborative instruction and learning. A survey of use of social network sites among Algerian students in the Institute of Library and Documentation Science at the University of Constantine 2 (Algeria) indicated a high use of social network sites among the surveyed population. Participants who filled in the survey questionnaire showed a preference for Facebook followed by LinkedIn presence and YouTube then twitter. Myspace does not seem to be so popular among Algerian students. Participants are using those social network sites mainly for communication and socializing. They expressed the view that social network sites can be used for academic purposes to discuss group project works, share assignments and course work, share files and lecture notes, create content, and exchange ideas. However, perceptions of respondents expressed through the survey do not indicate that the use of social network sites is likely to enhance academic achievements. Results of the study indicate that respondents are somehow apprehensive about social networks thinking that they will violate their privacy, are time consuming and discourage face-to-face communication. They also expressed concerns about quality control of information in social network sites.

Results of this study cannot be generalized as they reported the views expressed by the master students of Library and Documentation; only a small portion of students at

the University of Constantine 2 (Algeria). As the internet becomes more readily available on and off campus we will certainly see more students reverting to this new form of self-learning. We may also see in the near future faculty members using social network sites for formal education purposes as is the case in developed countries. Further research should consider investigating the use of social network sites by faculty members for formal academic activities and whether they perceive them as appropriate instruction and learning tools.

## References


1. Al-Muitairi, S. (2013). Use of Facebook by faculty in Kuwait University. Proceedings of the SLA/Gulf Chapter 19th conference on the future of library profession.

2. Battouche, K. (2012). Social networks and the globalization of culture: a study of Algerian youth institutions. A paper presented at the second international conference on "the place of social sciences and their role in studying contemporary issues: orientations and future prospects" Oman, December 16-18 (paper in Arabic).

3. Blake, Y, and Sikkal, J. (1996). Will the Internet revolutionize business education and research? *Sloan Management Review*, Vol. 37, No.3, pp.34-41.

4. Boumarafi, B. (2010). Strategies for the Delivery of e-information Services to Support the e-learning
Environment at the University of Sharjah. *The Electronic Library*, Vol.28, No 2, pp.276-285.

5. Boyd, D. M. and Ellison, N. B. (2008). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, Vol. 13, No.1, pp. 210–230.

6. Brotherton, P. (2011). Social network enhance employee learning. *ABI/IFORM Complete*, Vol. 65, No. 4, p.18.

7. comScore. (2007). Social networking goes global. Available at: http://www.comscore.com/press/release.asp?press=1555. Retrieved may 28, 2013

8. Dalsgaard, C. (2004). Social software: E-learning beyond learning management systems. Available at: http://www.imv.au.dk . Retrieved July 1st , 2013.

9. Downes, S. (2004). From Classrooms to Learning Environments: A Midrange Projection of E-Learning Technologies. *College Quarterly*, Volume 7, Number 3. http://www.collegequarterly.ca/2004-vol07-num03-summer/downes.html

10. Ellison, N. B.; Steinfield, C.; Lampe, C. (2007). The benefits of Facebook "friends:" social capital and college students' use of online social network sites. *Journal of Computer Mediated Communication*, Vol.12, No.4, pp. 1-26.

11. Friedman, L. W. and Friedman, H.H. (2008). High impact areas of the new media technologies: A review. Available at: http://www.morexpertise.com/download.php?id=97. Retrieved June 25, 2013.

12. Gentzler, A. and Oberhauser, A. (2011). College students' use of electronic communication with parents: links to loneliness, attachment and relationship quality. *Behavior and Social Networking,* Vol. 14, pp. 71-75.

13. Jabr, N. H. (2011). Social networking as a tool for extending academic learning and communication, *International Journal of Business and Social Science*, Vol.2, No. 12, pp.93- 102.

14. Krasnova, H. and Buxmann, P. (2012). Face book make users envious and dissatisfied, available at: http://www.hu-berlin.de/press-portal/nachrichten-en/nr1301/nr_130121_00
retrieved August 8, 2013.

15. Malesky, L.A. and Peters, C. (2012). Defining appropriate professional behavior for faculty and university students on social networking websites. *Higher Education*, Vol. 63, pp. 135-151.

16. Maney, K. (2009). "Next: An Internet Revolution in Higher Education." *Bloomberg Businessweek*. Available at: http://www.businessweek.com/technology/content/sep2009/tc20090914_969227.htm.
Retrieved July 13, 2013.

17. Reuben, R. (2008). The Use of Social network sites in Higher Education for Marketing and Communications: A Guide for Professionals in Higher Education. Available at: http://doteduguru.com/id423-social-media-uses-higher-educationmarketing-communication.html. Retrieved May 30, 2013.

18. Tiryakioglu, F. and Erzurum, F.(2011). Use of social networks as an educational use. *Contemporary Education Technology, Vol.* 2, No. 2, pp. 135-150.

19. Zuckerberg, M. (2011). Facebook. Available at: http://www.britannica.com/EBchecked/topic/1371433/Mark-Zuckerberg.
Retrieved july10, 2013

**bboumarafi@gmail.com**

iafor

# *The Importance of Process Mining in Enhancing Process Performance in Organisations*

Joos Meikhel Gaghenggang, Queensland University of Technology, Australia

The Asian Conference on Technology, Information & Society
Official Conference Proceedings 2014

**Abstract**
The development of business intelligence allows organisations to manage and enhance the decision-making process by providing methods and tools for analysing data. Process mining in organisation is needed to develop connection between data mining as business intelligence method and Business Process Management. The main purpose of process mining is to discover process model based on existing event log data that can be used for different objectives. This research will examine the essential concepts of process mining and its tools in order to analyse data and deliver proposal to enhance process performance in organisation. This research conducts focus group discussion as a qualitative method to discuss the advantages of process mining and to compare the process mining tools. The analysis highlights that the process mining has important role in organisation in determining: basic performance metrics; process model; and organisational model, and analysing social network and performance characteristics. Interestingly, both of process mining tools ProM and DISCO have different features and capabilities to discover the business process in organisation. This allows organisation to assess the data of business process transaction and provide some improvement approaches based on the result in process mining. By using the process-mining algorithm and tools, organisation can manage how to improve their process of business more effectively and efficiently in order to achieve their objectives.

Keywords: process mining, BPM, business intelligence, DISCO, PROM, data mining
Introduction

## Introduction

Dealing with business today, there are many data log and processes that perform an important role in assisting organisations to gain their business strategies by managing and analysing data log toward the achievement of desired business objectives. This experience requires a Business Process Analytics (BPA). BPA delivers business analyst, management, and stakeholders with understanding how to optimize business process in organisations. This concept consists a consideration either performance or compliance. In performance aspect, process analytics will help decision maker to shorten the time consuming in organisational process. On the other hand, in compliance aspect, process analytics will ensure that the process execution is align with rules and procedures (Muehlen & Shapiro, 2009).

The important purpose of this research is to examine essential concepts of process mining and its tools in order to analyse data and deliver proposal to enhance process performance in organisations. This includes gathering experience on using process mining tools, discussing the advantages of process mining and comparing the process mining tools.

## Research Approach

Qualitative research is used to conduct this research in order to collect data. Saldana (2009) states that primary data in qualitative research can be collected and analysed at least from interview, focus group discussion, documents, or website. This highlights that primary data can be gathered by using various techniques both direct and indirect. In addition, this method supports researchers to understand the phenomenon of research more clearly (Indulska, Hovorka, & Recker, 2010) This is important because this method allow the author to understand real condition or experience of using process mining concepts in order to explain how process mining can be used to improve business process in organisations.

This research conducts Focus Group Discussion (FGD) as a qualitative method to discuss the advantages of process mining and to compare the process mining tools. There are five groups involved in this research. Each group consist three to four students who were undertaking the Business Process Analytics Unit at QUT. The discussion is based on their experience on using process mining tools in the case study. Students worked with two different process mining tools: ProM and DISCO.

## Process Mining Tools

Process Mining is an approach to provide revolutionary concept to analyse and monitor the real life processes. Process Mining is focused on the extraction of knowledge about a (business) process from its process accomplishment logs (Prom, 2014). Basic requirement for process mining are:

a. Determine basic performance metrics, this element consist of two perspectives: process/control-flow and resources perspective.
b. Determine process model to discover a process model without knowing the business process.
c. Determine organizational model such as role and responsibilities and workload without knowing the organizational structure of organization.

d. Analyze social network based on handover of work, subcontracting, working together, reassignments, doing similar tasks
e. Analyze performance characteristics based on resources, performance metrics of case and machine learning techniques (van der Aalst, 2003)

This research will examine two process mining tools which is Prom and Disco. Both tools represent an extremely powerful data analysis tool that is capable of process discovery, process mining, and checking process conformance. Based on logs that are generated from a system, Prom or Disco is able to do multitudes of analysis. Both tools are extremely useful in performing resource and control-flow analysis that can be used for business process improvement initiatives, resource allocation, improved performance measures, relationship patterns and many more.

Based on the FGD, how to clean and filter the process log become major challenges. Business analyst should decide whether filtering out noisy data is necessary or not, because it would affect the process model. Moreover, filtered data have to be sufficient to provide justification and analysis to tackle the improvement process. The explanation of process mining tools is elaborated in following section:

**ProM**

Process Mining Framework (ProM) is an Open Source tools for process mining algorithm. It provides the framework that can be used and developed easily by users and developers. ProM is an extensible structure that provides a huge variety of process mining techniques in the method of plug-ins. This is an independent tool as it is implemented in Java, and can be downloaded for free.

Two versions of ProM were using in order to solve the improvement process in case study that is ProM 5.2 and ProM 6. The study of assessing the performance of process mining tools shows different experiences of using ProM 5.2 and ProM 6.3

The difference between both type of ProM as shown in the table below:

| ProM 5.2 | ProM 6.3 |
|---|---|
| Prom 5.2 is the preferred tool when doi | Prom 6 has better interface as compared |
| pecific analyses and conformance checks | Prom 5.2, although it is not as user-friendly |
| t has extensive analysis tools – resourc | ike Prom 5.2 it also contains an extensive l |
| ottlenecks, relationships, data perspectiv | f analysis tools |
| rganizational perspective, performance a | |
| nany more | |
| t conveys more data as compared to Prom | Multiple formulas can be applied in LT |
| nd is more user-friendly | hecker simultaneously, although the resu |
| | annot be verified using Disco and Prom 5.2 |
| | Social network model can be grouped |
| Suitable for analysts, but not for high-lev | Suitable for analysts, but not for high-lev |
| management | management |
| Process model has to be mined manually | Process model has to be mined manually |
| The tool is not compatible with Mac OS | The tool is not compatible with Mac OS |

Based on the findings this study highlight some aspects such as:

a. Serious user interface improvements are required in order to streamline and simplify the process of process mining with ProM.
b. The documentation for ProM and the Plug-ins are not good at all, there are so many broken links and half-finished documentation for the products.
c. ProM 5.2 is more efficient than ProM 6.3 as it lacks a wide range of plug-ins that the team could not find it capable to provide decent answers to each given task.
d. LTL Checker plug-in suffers from small issues that are slowing the steps of getting the results. One of the identified issues are, the user need to copy the needed term or resource from the system, which means typing it will not give a appropriate results, instead it will gives a results of 0.
e. Some of the LTL Checker formulas do not offer a clear description to its capabilities, which required the team to perform different inputs to understand its proficiency.

**DISCO**

DISCO is designed to make makes process mining easy and fast. Fuzzy miner is a basic miner in DISCO event tough it has been further developed in many approaches. The Fuzzy Miner was the initial mining algorithm to present the "map metaphor" to process mining, including advanced features like seamless process simplification and highlighting of frequent activities and paths (Gunther & Rozinat, 2010).

The study of assessing the case study using process miner, shows that there are some findings as explained below:

a. Charts and process models are automatically generated, but is limited to an EPC-like model
b. Animation of process model means that this is extremely useful for high-level management – easy to understand as animation shows process complexity, bottleneck, and reworks
c. Due to Disco simple use as well as answer can be found easily comparing to ProM, all group used Disco to compare and make sure about most of the tasks results.
d. Complex scenarios made simple using the filter, as compared to LTL checker
e. Using Disco to explore and identifying the solutions is easier than using ProM software, as the team found it difficult. It is because of its wide range of plug-ins that need first to be understood before applying to answer the task.

DISCO is powerful tool, because it shows several different visual data straightaway so business analyst can quickly have a look the duration or performance of processes. In addition, the process model of DISCO is automatically created that can give clear idea of process. Process animation helps business analyst to find out the potential bottleneck that can be reduced.

**Conclusion**

The analysis highlights that the process mining has important role in organisation in determining: basic performance metrics; process model; and organisational model, and analysing social network and performance characteristics. Interestingly, both of process mining tools ProM and DISCO have different features and capabilities to discover the

business process in organisation. This allows organisation to assess the data of business process transaction and provide some improvement approaches based on the result in process mining. By using the process-mining algorithm and tools, organisation can manage how to improve their process of business more effectively and efficiently in order to achieve their objectives.

## References

[1] Gunther, C., W & Rozinat, A. (2010). Disco: Discover Your Process. Retrieved from: http://ceur-ws.org/Vol-940/paper8.pdf

[2] Indulska, M., Hovorka, D., Recker, J. (2012). Quantitative Approaches to Content Analysis: Identifying Conceptual Drift Across Publication Outlets. *European Journal of Information Systems, 21 (1), pp. 49-69*.

[3] ProM. (2014). ProM Tools. Retrieved from http://www.promtools.org/doku.php

[4] Saldana, J. (2009). The coding manual for qualitative researchers. Los Angeles, CA: SAGE.

[5] van der Aalst, W. (2003). Process Mining: The next step in Business Process Management. Retrieved from https://www.google.com.au/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0CD8QFjAA&url=http%3A%2F%2Fwww.processmining.org%2F_media%2Fpresentations%2Fmining_au_process_days.ppt&ei=m2ajVIzrEM7UuQSEjIGQCg&usg=AFQjCNH6GO6ufBqPpSB7a10M_UkE8xcOig&bvm=bv.82001339,d.c2E

# *Enrich Cultural Heritage and Museum through Linked Open Data: The Content Analysis of Open Data Platform in Taiwan*

Shao-Chun Wu, National Palace Museum, Taiwan (R.O.C)

The Asian Conference on Technology, Information and Society 2014
Official Conference Proceedings

**Abstract**

Recently, the concept of Open Data has drawn attention worldwide. It is suggested that the government or government-subsidized institutions should disclose data for public access, distribution and reuse, therefore, those data can contribute more values to the society. In the trend of Open Data, museums or cultural institutions have started to release datasets over the internet, this is to invite people to use datasets of the museums, with the purpose of creating more software and services for the public. Consequently, the society will pay more attention to cultural institutions and facilitate a positive cycle. For instance, upon alliance, the museums in Europe founded the Open GLAM to implement the cultural Open Data project. In Japan, LODAC was constructed to implement Open Data in museums of Japan. The government of Taiwan established the Open Data Platform last year. However, the characteristics of datasets of museums on the platform were not systematically explored. Therefore, this study treats Open Data Platform of Taiwanese government as the subject, use SPSS to conducts content analysis to probe into 186 datasets related to museums or cultural institutions. The purposes are to analyze the data content, data format and characteristics, as well as to derive means to enhance data usage. Finally, suggestions are proposed for future implementation of Open Data in museums of Taiwan. Thus, museums' data in Taiwan can be used by the public through continuous open data, and in turn enriching cultural and educational functions of museums.

Keywords: Open Data, Museum, Taiwan, Cultural Heritage, Linked Open Data (LOD)

iafor

The International Academic Forum
www.iafor.org

## Introduction

The term “open data” refers to data disclosed to the public by government agencies (or private agencies that have received government subsidies) via the Internet. These open data can be accessed, spread, and reused by the public; they may even be used for commercial purposes (Baltussen, Oomen, Brinkerink, Zeinstra & Timmermans, 2013; Janssen, Charalabidis & Zuiderwijk, 2012). To promote and advocate open data in recent years, a substantial amount of data has been released by governments around the world; these data are used to develop numerous creative applications. In addition, by integrating open data with “linked data,” a method that is based on the Semantic Web technology, “linked open data” (LOD) can be formed. These endeavors have attracted considerable attention of the world to the subject of “open data” as well as the future applications and developments of these data.

The release of open data by government departments originated in 2009, when the U.S. government built data.gov to provide the U.S. Government’s open data. In 2010, the United Kingdom Government followed suit and created data.gov.uk (Chang, 2012). With the LOD concept gaining prominence and linked data showing great development potential as an artifact-searching tool between different domains, museums and cultural institutions began to actively promote open data and LOD (Van Hooland, Verborgh, Wilde, Hercher, Mannens & Walle, 2012). Similarly, in 2012, the Taiwanese government built an inter-government department open data platform that disclosed datasets of the government departments for the public to access. These datasets included those from museums as well as cultural institutions.

Although a number of museum-related datasets have already been released on Taiwanese government’s open data platforms, systematic discussions on the characteristics that these datasets and how these datasets can be effectively used to enhance the value of museums remain absent. In addition, studies on museum-related datasets and the raising of social values of cultural institutions remain deficient. Therefore, in this study, we will use 186 museum-related datasets from the Taiwanese government’s open data platform as the study subjects. Descriptive statistics will be performed using the content analysis method and SPSS Statistics 21.0, which will facilitate the investigation of the characteristics that these datasets have and to make recommendations to the museum industry concerning LOD development. The objective is for the museum industry of Taiwan to improve its values through the use of open data and LOD.

## Background

### Development history of open data and LOD

Although the concept of open data apparently originated in 2009, the development history of open data can be dated back decades ago. Since the 1990s, governments around the world had been actively promoting eGovernment with the anticipation of improving government efficiency and government-citizen interactions (Chen and Liao, 2013). Following the rapid spread of the Internet and the government's emphasis of public participation in recent years, open data, which featured public value creation characteristics, began to emerge in eGovernment environments. However, differences exist between open data and eGovernment; the biggest of which is that eGovernment emphasizes the ability of the public to access government data or services via the Internet, whereas open data considers the ability of the public to reuse or even make profits from using the data disclosed by government agencies (or private agencies that have received government subsidies). In essence, the concept of open data is built on two main premises: First, the public has already developed the ability to interpret and use source data. The government and the social groups have also learned the ability to think, discuss, and exchange views openly; second, open data can be used to create a more open government (Janssen, Charalabidis & Zuiderwijk, 2012).

In general, to the public, the greatest advantage of open data is gained when all sectors of the society fully utilize the data released by the government to create services needed by recipients such as the public, tourists, and businesses. Therefore, as a whole, source data disclosed on open data platforms promoted by governments around the world are displayed in the form of datasets so that the public can download them; sometimes, even open APIs are provided (Chang, 2011). Open data promoted by governments also have the effect of improving citizen participation and information dissemination (Chen & Liao, 2012). Others argue that open data feature additional advantages: socially and politically, they enhance the transparency of government operations, elevate public satisfaction, fuel the development of knowledge, and improve the government's decision-making process; economically, they stimulate competition and increase the likelihood of new product and service creations; technically and business operation-wise, they optimize administrative operations and increase the likelihood for the public sector to resolve issues using external mechanisms (Janssen, Charalabidis & Zuiderwijk, 2012). However, the promotion of open data by government departments is not without obstacles or side effects and support from social environments is necessary for the promotion of open

data. In general, technical obstacles encountered during the promotion of open data include difficulty providing, accessing, searching, and using data. Problems such as poor data quality, unclear metadata, and unfavorable data compatibility or linkage also exist (Zuiderwijk, Janssen, Choenni, Meijer & Alibaks, 2012). The failure of the government agencies to introduce relevant policies, complementary laws, and appropriate revenue-generating mechanisms also increase the number of obstacles involved with the promotion of open data (Janssen, Charalabidis & Zuiderwijk, 2012).

In recent years, open data are integrated with Semantic Web technology to create LOD, which improves the linkage between data and enables the reading of online information (most of which is found on webpages) to evolve from human-only to machines. Current Semantic Web technology primarily involves the use of resource description framework (RDF) to facilitate data description, uniform resource identifier (URI) to enable data display, and standard SPARQL to assist data query. In addition, Semantic Web technology allows data to be linked to other URIs or data hubs (Bizer, Heath & Berners-Lee, 2011).

## Current Promotion of LOD in Museums

### Museums and LOD

The aforementioned development history of open data and LOD shows that open data have become an important method for promoting public participation and improving the effectiveness of information dissemination. Because museums are accessible to the public, can be considered a form of educational institution, and receive subsidies and donations from the government and the public, respectively, many museums join the ranks as providers of open data; some even become promoters of LOD. In fact, by publicizing existing digital data, museums not only show that their cultural artifacts can be used to innovate and drive new artifact developments, but they also fulfill their missions to serve and educate the public. These efforts not only allow museum materials to serve their purpose inside the museums, but also enable them to be used by the public outside the museums, achieving the goals of museum education, museum promotion, research support as well as the promotion of learning (Baltussen, Oomen, Brinkerink, Zeinstra & Timmermans, 2013). Therefore, many museums home and abroad engage in the use of not only open data, but also LOD with the aim of not only taking full advantage of open data but also connecting it to Semantic Web technology. This enables data of different formats and hardware environments used in different museums to come together on the Internet and allows the reading of such

data to be done by machines. Data from different museums can thus be accessed, which maximizes the benefits of museum data.

Specifically, the promotion of LOD offers the following advantages:

(1) Promotes the use of open data, particularly the disclosure of source data;
(2) Enhances the richness of cultural data by linking the data from different museums, enabling scholars to conduct studies that were previously difficult (Hsiao, 2013);
(3) Facilitates the design and development of terminal interfaces for Internet users, providing users with unprecedented experience (Hsiao, 2013); and
(4) Improves integrated data query and the effectiveness of data exchange.

To achieve the aforementioned objectives, museums must use the standard Semantic Web technology or whichever Semantic Web technology made available to them to connect or produce data. Current attempts by museums to promote LOD generally emulate the methods adopted for Semantic Web promotion. These methods include RDF and SPARQL; currently, a number of software tools are available for converting database data into the RDF format (Szekely, Knoblock & Wang, 2013). However, for ontology, specific standards (e.g., the CIDOC Conceptual Reference Model (CIDOC-CRM)) used by museum industries must be followed. In addition, data vocabularies must be modified to match those used by other museums. In general, for museums to develop LOD, they must first convert existing museum data into RDF data, modify the vocabularies, and connect the generated data with external data hubs. However, when promoting LOD, museums face difficulties such as shortages of manpower and funding, lack of standards controlling data content and vocabularies, and problems converting museum database into RDF data because of overly complex database structure and content (Szekely, Knoblock & Wang, 2013).

**The promotion of open data and LOD by museums as a unit**

Because museum industries around the world are beginning to notice the tremendous effect of LOD promotion on artifact marketing, which elevates the values of the artifacts, museum-related professional organizations are starting to encourage museums to promote LOD. Some museums even worked together to develop LOD. For example, the United Kingdom's Open Knowledge Foundation founded a cooperative project called "Open Galleries, Libraries, Archives and Museums" (OpenGLAM), which called together museums and cultural institutions to host workshops as well as technology and information exchanges to promote the use of LOD by museums (Hsiao, 2013). LODLAM, another professional organization, has

similar missions.

In addition to the efforts made by OpenGLAM to promote LOD, many museums supported LOD by using them in practice. Beginning from 2012, European museums founded the European Project, providing descriptions of artifacts from over 1,500 European cultural institutions using a uniform data model called the European Data Model (EDM). These artifact descriptions are published using the RDF model and released using the CC (Haslhofer & Isaac, 2011). Apart from the European Project that promotes the LOD by having museums from different countries working together, some museums team up with domestic museums to promote LOD and open data. For instance, in 2011, the Netherlands initiated the Open Culture Data project, gathering six of the Netherlands' cultural institutions to promote open data and form a tight network for them to exchange ideas, offer courses to share experiences, and discuss various technical and legal issues (Baltussen, Oomen, Brinkerink, Zeinstra & Timmermans, 2013). In 2010, Japan launched the Linked Open Data for Academia (LODAC) project, bringing together 15 of the museums in Japan and providing them with the appropriate data motel to enable them to publish the RDF data and to connect to the data hub (Kamura & Takeda, 2013). Fig. 1 shows the data query interface used for the European project.

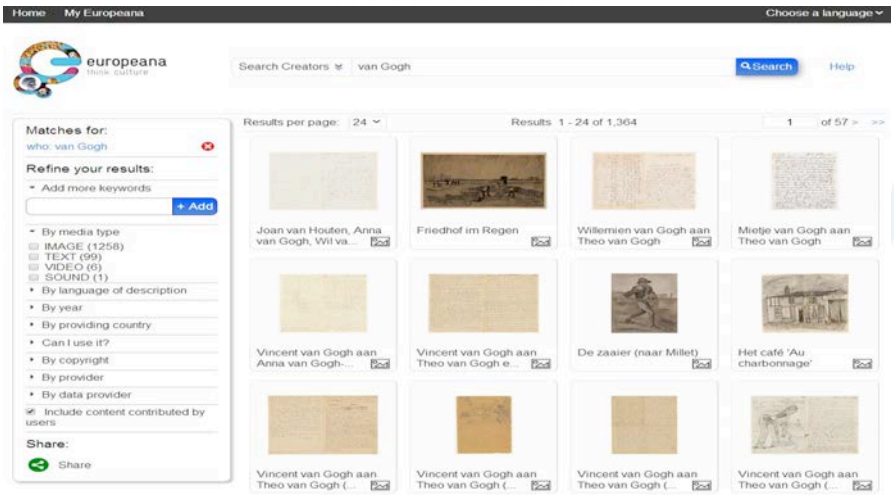



Figure 1. The Data Query Interface of Eurpean Project

**Efforts made by each individual museum**

In addition to collaborating with other museums to develop open data and LOD, each individual museum is able to, within the scope of their ability, promote open data or develop LOD. For example, since a few years ago, the British Museum used the CIDOC-CRM to describe the properties and relationships between data fields for the

field of ontology. The data were published in RDF format and the SPARQL Query service was provided. The artifact data of the British Museum, in RDF format, can also be downloaded from its official website. Moreover, the Powerhouse Museum (in Australia), the Smithsonian American Art Museum (SAAM; in the U.S.), and the Cooper-Hewitt National Design Museum (also in the U.S.) compare, process, and publish artifact metadata using various Semantic Web technology to enable them to be queried and used by the public (Van Hooland & Verborgn, 2014; Szekely, Knoblock, Yang, Zhu, Fink, Allen & Goodlander, 2013; Hooland, Verborgh, Wilde, Hercher, Mannens & de Walle, 2013).

**Taiwanese government's open data platform and museums' datasets**

To follow the global trend of open data promotion, in 2012, the Taiwanese government established its own open data platform and demanded that all departments of the public sector, including museums, to upload a set amount of datasets to the platform. The public can query or browse through the datasets provided by each government department by theme, name of the organization, or other categories. However, the category "museum" is yet to be made available. Therefore, users who wish to query or download museum datasets must try different query categories to find museum-related datasets. Currently, Taiwanese researchers are attempting to group culture-related datasets into a separate category on the said platform and are experimenting the feasibility of establishing a data model that can be used by the cultural industry (Chen & Liao, 2013). The goal is to use it as the basis for developing LOD for Taiwanese museums in the future. Fig. 2 shows the user interface of the Taiwanese government's open data platform.


Figure 2. The Interface of the Taiwanese Government (R.O.C)'s Open Data Platform

Taiwan's central government also encourages local departments to build their own open data platform on their official website and to provide services that can facilitate data circulation and usage. Concerning the cultural industry, departments such as the National Palace Museum and the Ministry of Culture have all established their own open data platform; the Taipei City Government's open data platform also provides various culture-related LOD services (Wu, Yang & Tsui, 2012).

## Research Method

In this study, museum-related datasets obtained from the open data platform of Taiwanese local governments will be selected as the study subjects. The content analysis method will be employed as the research method. The content analysis method is primarily used to analyze the content of specific data; statistical methods and other methods are used to understand the hidden characteristics of the content, such as newspaper content and TV program dialogues. The content analysis method has been commonly used in the field of mass communication (Busha & Harter, 1980); some had also used this method to study website content (Wu, 2002). The said method was used to analyze the characteristics of 186 museum-related datasets (e.g., content, format, and publishing departments) on the Taiwanese government's open data platform, and SPSS version 21.0 was used to perform descriptive statistics to identify the characteristics of these datasets. All data collection was completed by Sept. 2014.

## Research Results

### Number of Museum-Related Datasets on the Platform

Because the category "museum" remains to be found on the Taiwanese government's open data platform, to find out the number of museum-related datasets on the platform, users must search for and browse through the datasets from various categories. By adopting this method, a total of 186 museum-related datasets were identified by Sept. 2014. These datasets were scattered throughout the various categories such as cultural events and tourism; some were located by searching under the institutions responsible for managing museums. No museum-related datasets were found by simply entering museum-related datasets in the search bar. Further analysis showed that most of the museum-related datasets published on the platform were 3 star structured and unstructured data and that the data model used by museums to create the datasets

differed. In addition, the following results were obtained: the specialized vocabularies used by the museums differed; the LOD failed to match the 5 star LOD standards put forward by Tim Berners-Lee; no link nor association were found between the datasets; and no common API services that were shared across museums were found.

**Institutions that provided museum-related datasets**

There are three institutions that currently provide museum-related datasets on the aforementioned open data platform. They are the National Palace Museum, the Ministry of Education (that oversees nearly 10 public museums), and the Ministry of Culture (that oversees tens of public museums and cultural institutions as well as grants subsidies to private museums; the said public museums and cultural institutions all vary in size). The datasets provided by the three institutions differed; the number and characteristics of these datasets also differed. Concerning the National Palace Museum, it had uploaded 50 datasets to the platform for users to download. The datasets comprised exclusively metadata of the artifacts, which were their "source data." Regarding the museums overseen by the Ministry of Education, they had uploaded 21 datasets to the platform; the datasets contained zero artifact metadata. With respect to the museums and cultural institutions overseen by the Ministry of Culture, they had uploaded 115 datasets to the platform; the datasets contained artifact metadata as well as other information. Although the datasets primarily came from public museums, those that were provided by private museums were also found, albeit more difficult to find. Fig. 3 shows a diagram comparing the number of museum-related datasets provided by the three institutions.

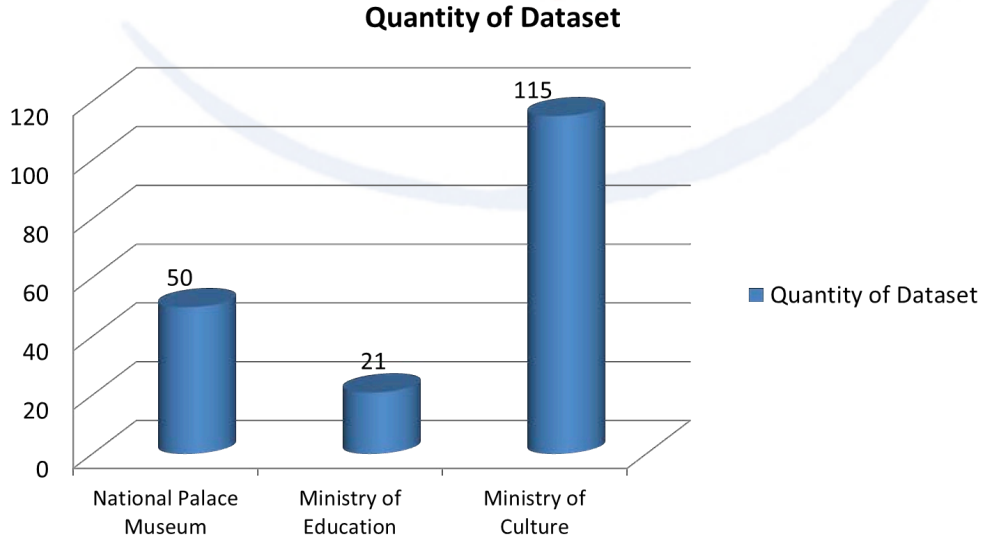



Figure 3. Number of Museum-related Datasets Provided by Three Institutions

**Types of content found in museum-related datasets**

Museums are a place that can be accessed by the public and one that can be used to educate. Museums provide source data of artifacts as well as other data to the public to enable it to create museum and culture-related software and services. Current datasets uploaded by museums to the open data platform feature a wide variety of content. Based on the characteristics of these contents, they can be divided into museum-related activities, source data of artifacts, descriptions of permanent and special exhibitions, museum's public facilities and cultural parks, learning resources, and others. A analysis of the 186 datasets showed the number of datasets for the six categories, which are listed as follows: museum-related activities, 17; source data of artifacts, 149; descriptions of permanent and special exhibitions, 3; museum's public facilities and cultural parks, 13; learning resources, 3; and others, 1. Fig. 4 provides a schematic diagram of the dataset content breakdown (in percentages) .

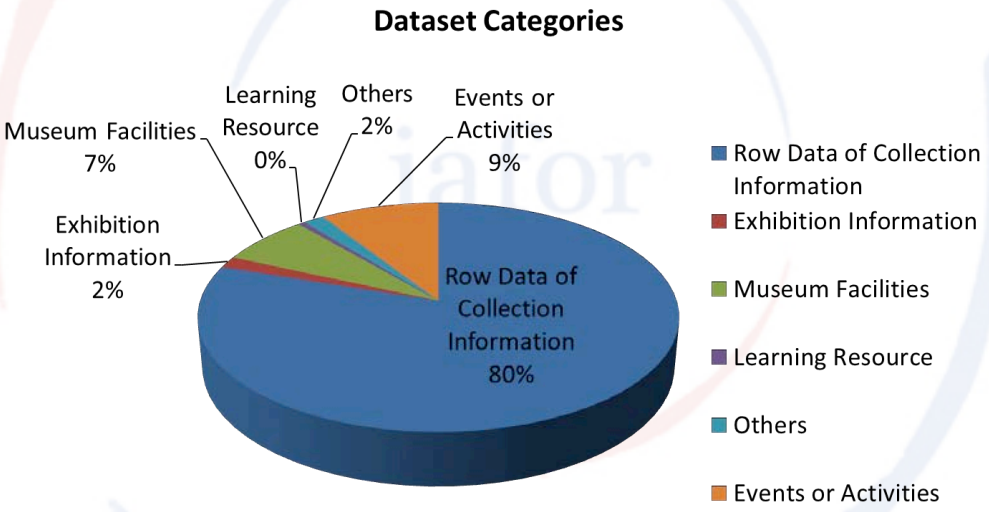



Figure 4. Content of Museum-related Datasets (in percentages)

A breakdown of the museums' dataset content reveals that artifacts remain the core asset of museums; they are also museums' most unique and irreplaceable data. A total of 149 datasets were source data of artifacts; although museums are responsible for preserving artifacts, their most important missions are to promote artifacts and educate the public. Therefore, museums have the tradition of designing numerous educational activities and teaching materials. However, of the datasets uploaded to the platform, learning resources are noticeably scant, which warrants further attention.

**Format of museum-related dataset**

The 149 museum-related datasets featured a variety of formats. Because the dataset formats differed between the government departments, it showed that the government departments had different views regarding the use of Semantic Web in the future development of museums as well as the use of planned and large-scale open data promotions. Of the 149 datasets, 62 were in xml format, 81 were in JSON format, and 9 were in CSV format; 18 were in both XML and JSON formats; and 16 were in XML, JSON, and Excel format. Fig. 5 shows a breakdown of the dataset types (in percentage) for each format.

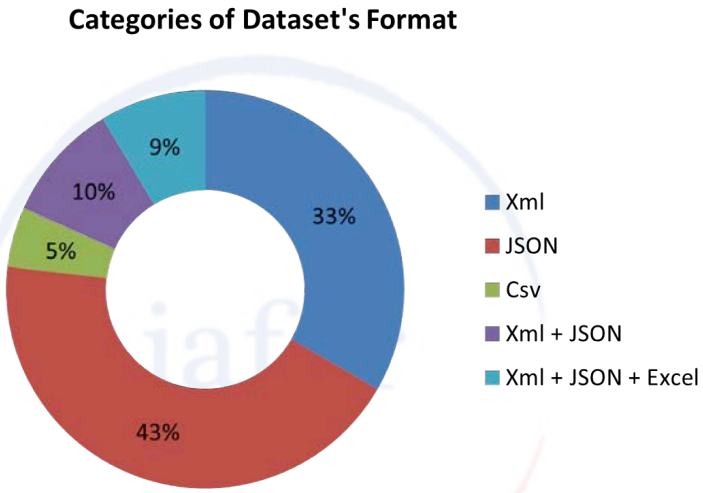



Figure 5. Formats of Museum-related Datasets (in percentage)

Among the various datasets types, those that were in XML or JSON format accounted for the largest proportion. These datasets were primarily provided by the National Palace Museum and the Ministry of Culture, which also provided the source data of museum artifacts. Datasets in CSV format were mostly provided by museums overseen by the Ministry of Education and no source data of artifacts were found. These results indicate that the National Palace Museum and the Ministry of Culture are more concerned with the development of open data; their plans are also more thorough.

**Providing institutions with complementary measures and promotional activities to market museum-related datasets**

The Taiwanese government's open data platform is used by various government departments (including public museums) to upload datasets to the platform for the

public to download. However, in order for museums' open data to be fully used by the society and to promote the Semantic Web technology-based LOD, museums must introduce complementary measures and promotional activities. Nevertheless, the complementary measures and promotional activities put forth by departments that release museum-related datasets on the Taiwanese government's open data platform are high dissimilar. A detail of the current situations is provided as follows:

1. Only the National Palace Museum and the Ministry of Culture have developed a special webpage on their official website, providing open data-related services;
2. Only the National Palace Museum has engaged in large scale open data promotional activities such as app competitions and seminars;
3. Only the Ministry of Culture provides API services and has formulated policies concerning the use of and applications for API; and
4. Only the Ministry of Culture and the National Palace Museum have issued the guidelines regarding the use and authorization of open data; authorization-related policies remain missing in most other museums.

**Conclusion**

The aforementioned analysis of the Taiwanese government's open data platform and statistical analyses of the datasets showed and compared the efforts made by foreign museums to promote LOD and open data with those by the domestic museum industry. The following observations were made:

(1) The Taiwanese government's open data platform is the biggest platform for publishing datasets. However, the category "museum" remains to be seen in the open data platform. In addition, searching for museum-related datasets by entering museum-related datasets in the search bar produces no results;

(2) Most of the datasets released by museums on the Taiwanese government's open data platform are textual and/or numerical in nature; image, video, and audio data remain deficient. Although there are quite a few museum artifact datasets, the majority of these datasets are artifact metadata and artifact-related images or multimedia materials are rarely categorized as open data. With respect to the number of datasets, there is still a lot of room for improvement to be made by Taiwanese museums to increase dataset numbers as well as in introducing more diverse dataset formats and content to facilitate the use of museum data. However, for more museum data or images to be included as a part of open data, museums must be aware of issues

such as intellectual property rights and management as well as related authorization policies. These issues all require the museums' due diligence to ensure a balance between museum management, profit creation, and the provision of accessible open data to the public;

(3) In a more strict sense, compared with the datasets provided by the British Museum and the museums that founded the European project, linked data provided by domestic institutions are not high in quality. Museums are suggested to work together to ensure that the use of vocabularies and the development of ontology concepts are consistent and that mature Semantic Web standards are employed to facilitate museums' development of LOD in Taiwan;

(4) Compared with the efforts to promote LOD in Europe, those made in Taiwan primarily come from public museums; efforts made by private museums remain insignificant. Seeing how public museums in Europe work together to promote open data and LOD, it brings into question whether private museums in Taiwan, especially those that receive government subsidies, should work jointly to promote open data and whether appropriate strategies should be devised; and

(5) To improve the level of consistency of the standards and the linkage between the museum datasets, the museum industries in Europe and Japan generally work together to promote open data or LOD. The museums also cooperate with each other and engage in various types of promotional activities. Similar methodology should be emulated by the museum industry in Taiwan; comprehensive planning or a collaborative approach should be used to promote open data and LOD. Various types of promotional activities may be hosted to elevate the level of consistency of museum data and improve the public's understanding of such data.

The results from the above analysis of Taiwanese government's open data platform and the statistical breakdown of the datasets were used to examine the efforts made by the museum industry in Taiwan to promote LOD and open data; these endeavors were compared with those made by foreign museums, which reveals that there remains a lot of room for improvement for the museum industry in Taiwan.

In short, the focus of Taiwan's museum industry in the future should be on improving the current deficiencies in LOD promotion so that museum data can be used by the public. This ensures the continued circulation and use of museum data in the information society, enabling the data to be utilized not only internally for museum

management, but also externally by the public for education, research, learning, and cultural and creative works. This will facilitate the public's understanding of museums and elevate their support of museums.

## References

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems, 5*(3), 1-22.

Baltussen, L. B., Oomen, J., Brinkerink, M., Zeinstra, M., & Timmermans, M. (2013). Open Culture Data: Opening GLAM Data Bottom-up. In N. Proctor, & R. Cherry(Eds.), Museums and the Web 2013: Selected Papers from an International Conference (pp. 61-78).

Busha, C.H. & Harter, S.P. (1980). *Research Methods in Librarianship : Techniques and Interpretation.* New York : Academic Press.

Chang, Jia-Sheng (2012). The open data and Taipei government. *Government Resource Planning, 36*(4), 61-69。

Chen, Zhun-Mei & Liao, Jun-Feng (2013). *Design and Implementation of Linked Open Government Data Service for Culture Domain.* Retrieved December 29, 2014, from http://tcrc.tanet.edu.tw/TANET2013/paper/G11-750-2.pdf

De Boer, V., Wielemaker, J., Van Gent, J., Hildebrand, M., Isaac, A., Van Ossenbruggen, J., & Schreiber, G. (2012). Supporting linked data production for cultural heritage institutes: the amsterdam museum case study. In The Semantic Web: Research and Applications (pp. 733-747). Springer Berlin Heidelberg.

Haslhofer, B., & Isaac, A. (2011, September). data. europeana. eu: The europeana linked open data pilot. In *International Conference on Dublin Core and Metadata Applications* (pp. 94-104).

Hsiao, Ching-teng (2013). Linked Open data and its Application in Museums. *The Newsletter of the Chinese Association of Museums, 64*, 15-18.

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29, 258-268.

Kamura, T., Takeda, H., Ohmukai, I., Kato, F., Takahashi, T., & Ueda, H. (2011). Study support and integration of cultural information resources with linked data. In

Culture and Computing (Culture Computing), 2011 Second International Conference on (pp. 177-178). IEEE.

Szekely, P., Knoblock, C. A., Yang, F., Zhu, X., Fink, E. E., Allen, R., & Goodlander, G. (2013). Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In *The Semantic Web: Semantics and Big Data* (pp. 593-607). Springer Berlin Heidelberg.

Szekely, P., Knoblock, C., & Wan, J. Karma: Tools for Mapping Collection Meta-Data
to Linked Open Data. In Museums and the Web 2014, N. Proctor & R. Cherry (eds). Silver Spring, MD: Museums and the Web. Published February 1, 2014. Consulted October 4, 2014. http://mw2014.museumsandtheweb.com/paper/karma-tools-for-mapping-collection-meta-data-to-linked-open-data/

Van Hooland, S., Verborgh, R., De Wilde, M., Hercher, J., Mannens, E., & Van De Walle, R. (2012, May). Evaluating the success of vocabulary reconciliation for cultural heritage collections. *Journal of the American Society for Information Science and Technology*, 64(3), 464-479.

Van Hooland, S. & Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums*. London : Facet Publishing.

Wu, Shao-Chun (2002). Content Analysis Method and Its Application in Librarianship
Studies. *Journal of Librarianship and Information Studies*, *40*, 47-61.

Wu, T. T., Yang, W. H., & Tsui, W. (2012). Technologies and Applications of Semantic Web, Linked Data and Open Data. *ICL Technical Journal, 145*, 102-109.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Sheikh_Alibaks, R. (2012). Socio-technical impediments of open data. *Electronic Journal of e-Government*, *10*(2), 156-172.

# *Creating A Social Media Based Continuous Interaction Platform For The Design Studio*

Kutay Guler, Dumlupinar University, Turkey

The Asian Conference on Technology, Information & Society
Official Conference Proceedings 2014

**Abstract**
Social Media offers enticing pedagogic opportunities. Unlike the users of Learning Management Systems (LMSs) or Personal Learning Environments (PLEs), the Social Media user has the ability to create a profile and form interactions with his/her contacts which have real life value and impact. In a learning context, considering the popularity of Social Media, user familiarity with the interface and the workflow can also be advantageous. A possible pedagogic implementation of social media could be a continuous interaction platform for the contemporary design studio which might help improve issues such as; efficiency of studio critiques, peer interaction, and homogeneous distribution of course load throughout the semester.

In this study the effectiveness of traditional and social media based design studio courses are investigated and compared. The sample population involved 78 4th year Graphic Design students enrolled in "Exhibiting and Exhibition Graphics" course, which was divided into a treatment and a control group each consisting 39 individuals. For the treatment group, the studio process was performed solely on social media and all interactions were kept strictly online. For the control group a standard studio process was employed. The evaluation of the aforementioned two processes was made both from the students' point of view – where students' opinions and experiences were compared, and from the instructors' point of view – where a panel evaluation of the submitted works were compared. According to findings the implementation of social media in design studios and the future of the concept are discussed.
*Keywords* – Social Media, Design Studio, Studio Critiques, Web 2.0, Facebook

## Introduction

Social media is usually defined through the ideological and technological foundation introduced by Web 2.0. Contrary to the static, non-participatory and limited environment of Web 1.0, Web 2.0 provided an evolving, participatory and content-rich platform (Goodchild, 2007; O'Reilly, 2005; 2007; Özata, 2013; Singh, 2010; Stark, 2006). The Bulletin Board Systems from the 1980s and instant messaging services of the previous era, such as IRC, ICQ and Microsoft Messenger are considered as the precursors to social media and networking (Özünlü, 2012). Compared to the previous installment, Web 2.0 users aren't passive receptors of information anymore but are an active part of content creation. The content can be accessed independent of time and place, can be shared with larger audiences and it can be enriched with contribution from other users (Anklam, 2009; Boyd, 2012; Komito and Bates, 2009; McLoughlin and Lee, 2007; Robinson, 2007).Some prominent features of Social Media are: (1)the user can create a personal profile that is relatable for other users, (2) the user can form connections through following and befriending other users while starting open or private conversations, (3) the content can be shared, searched and referred back to independent of time and place(Boyd and Ellison, 2007; Conole and Culver, 2010).Additionally, efficient and effective communication features, satisfying curiosity about others, gaining popularity and forming/improving new relationships are listed as reasons for the prevalence of social media use amongst the young (Urista, Dong and Day, 2009).

Social media has become a widely researched subject especially in the field of education, as it is ubiquitous, accessible, responsive and continuously evolving. Academic studies on social media often focus on creating profiles, sharing material, leaving comments and forming friend/contact lists (Mazmani and Usluel, 2010).This study seeks to contribute to the growing body of research by proposing a framework for implementing and evaluating social media in design studio learning. The differences between the traditional design studio learning and a social media based design studio learning are investigated by contrasting student performances and opinions. Through the analysis of data, this study aims to determine whether the quality of learning and social interaction in the design studio environment changes depending on the learning platform. This study will also provide an initial groundwork for future studies, especially in the field of design.

## Design Studio and Online Education

Design studio is a process during which learning occurs while all the skills and knowledge acquired by the student is focused on solving a design problem over a period of time. Design studios are particularly important in design education as they provide a simulation of the actual professional environment (Brusasco, 2000). The learning process in the design studio can be defined as a series of cyclical iterations based on a continuous dialogue between the instructor and the student, through which the given design problem is solved (Schön, 1987; Kurt, 2009)(Figure 1). Raw information is rarely provided during the design studio process, rather the student is expected to generate knowledge with the guidance of the instructor. This statement implies that the course of creating a solution is as important as the final product.

Critiques form the foundation of learning in the design studio. Critiques provide instructor and peer input that will improve the design solution that is being developed by the student. Critiques will lead to revisions, many times during the course of the design studio. There are several ways to receive critiques. Besides desk critiques with the instructor, students can also receive critiques during presentations, pin-ups, midterm juries, and final juries from their peers and guest designers/instructors. Usually the process is open and social, students can spectate their peers during critiques and assess their own progress by comparison.

The widely adopted traditional design studio mechanic has several well-documented shortcomings. Usually social interaction is limited to class hours during which the instructor is expected to provide desk critiques for 12 students in one 4 hour session on average (Tate and Osborne, 2013). Cochrane and Bateman (2010) report that students often fail to track and document their progress and have difficulty in remembering how they achieved a particular solution. This is due to the fact that the concept of design studio is foreign to the students whom haven't encountered a similar learning process before. As a result adaptation problems sometimes even extend to final years of education. Besides inadequate class hours and adaptation problems, workload distribution is also skewed. Kurt (2009) reports that in a typical traditional design studio usually 80% of the work is finished during the 10% of the duration of the course. Lastly presentations, pin-ups, juries and even desk critiques might become emotionally overwhelming and affect learning negatively as factors such as shame and fear come into the equation in such social circumstances (Utaberta, et al., 2011).

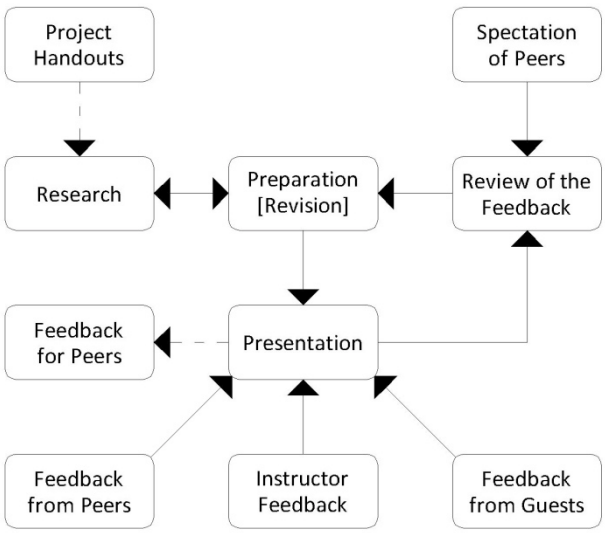



**Figure 1.** Interactions in a design studio environment.

Computer mediated communication stands out as an alternative to the social interactions of the traditional design studio environment. The positive and negative aspects of computer mediated communication has been scrutinized in numerous academic research (Rheingold, 1994; Jones, 1997; Robins and Webster, 1999). It is often argued that virtual communication experience is a simplified version of real-life social interaction. However, virtual communication experience do have real-life propagation and social consequences (Bayraktutan, 2013). In design studio context it can be argued that face to face social interaction during presentations, pin-ups,

critiques and juries is a more ideal situation compared to virtual communication experience where signifiers such as facial expressions, body language, tone of voice or the rhythm of speech are missing. Even though video cameras and conferencing options may be used as supporting tools, the quality of communication during a real-life interaction is hard to achieve (Hoyt, 2000). However the quality communication is a trade off for a communication environment without the limitations of time and space. Martin (2001) argues that subjects starts to compensate for the missing traits of communication by being more descriptive and actively seeking understanding. It can be argued that in a learning environment such behavior might be a positive. Cochrane and Bateman (2010) claim that increased amount of communication environment that is not bound by time and space limitations creates a unique situation for an improved learning environment. Virtual communication experience also creates an opportunity for mobility.

Considering the shortcomings of the traditional design studio environment and possible benefits of a computer mediated alternative, a continuous interaction platform is proposed in this study. The ability to provide unlimited communication and archival capabilities would be the best option for a host (Dickey, 2007; Jonassen, 1999). Social networking services proved to be the better host for the proposed continuous interaction platform as they provided a familiar interface for users, frequency of visits and interactions, profile creation and real-life social impact, easy system management and minimal downtime. Learning Management Systems (LMS) such as Adobe Connect, Blackboard or Moodle weren't considered as they were expensive, hard to manage/maintain, did not provide similar communication possibilities as social networking services, which was of utmost importance (Tu and Isaacs, 2000). Additionally form LMSs, limited access and limited mobility, unfamiliarity with the interface and dynamics of utilization created a risk of affecting learner performance negatively (Uden, 2007; McLaughlin et al., 2008).

In order to implement the continuous interaction platform, a private group was created on the social networking service Facebook. Among all social networking services Facebook was chosen as it was widespread, familiar, functional and reliable; meeting the previously mentioned host requirements. The prevalence and familiarity of the host system shortens the time period spent on orientation and adaptation. Additionally the Facebook app on smart phones enables the user to have truly continuous social interaction, as the user gets a continuous stream of notifications. Besides its advantages as the host, one major limitation is Facebook isn't designed for any educational process in mind and does not provide tools that focus on learning. Additionally, there is a continuous stream of events, media items, suggestions and advertisement acting as distractions. As Kaplan and Haenlein (2010) emphasized it was necessary to create guidelines and limitations while working with social media.

## The Experiment and Procedure

In order to evaluate the proposed continuous interaction platform an experiment was devised involving 75 fourth year graphic design students. The sample population was divided into a treatment group of 40 students who experienced design studio on the continuous interaction platform and a control group of 35 students who experienced a traditional design studio environment. Due to very specific sampling requirements, purposive sampling was used for this qualitative study (Given 2008; Salkind 2006).

Sample population's grade point average (GPA) indicated that, members of both groups have achieved equivalent academic success on average. The sample population was required to be able to understand basic spatial relationships, basic ergonomics and basic visitor dynamics in order to solve the design problem presented in the experiment. These basic skills were acquired by the sample population during the elective Space and Exhibition Design course in the previous semester. Besides the general requirements the treatment group was required to have a familiarity as a user with the social networking service Facebook. Both groups were expected to receive critiques, share course related material, and actively participate in the ongoing discussion in their respective learning environments. Students were expected to receive at least three critiques throughout the duration of the course and were made aware of the fact that their cumulative contribution would affect their grading.

The design problem to be solved was designing an advertising and distribution stand for a selected periodical on a 5.25 x 3.0 meter area with a possible depth of 0.75 meters (Figure 2). There was a 2.0 x 3.0 x 0.5 meter niche on the surface of the given geometry which the student could move around, divide, and re-shape while keeping the ratio between positive and negative space. Students were expected to submit a single A2 paper at the end of the course showing a front view, a perspective view, measurements, and notes on concept and functionality. Students were given a template for the submission to minimize missing information as the evaluators of the project would be foreign to individual projects. Each submission was evaluated by three evaluators, all of whom had previously taught and practiced exhibition design. The evaluators were requested to grade the submissions according to quality of conceptualization/form, functionality/ergonomics, originality/creativity, and visuals/presentation on a 10 point scale.

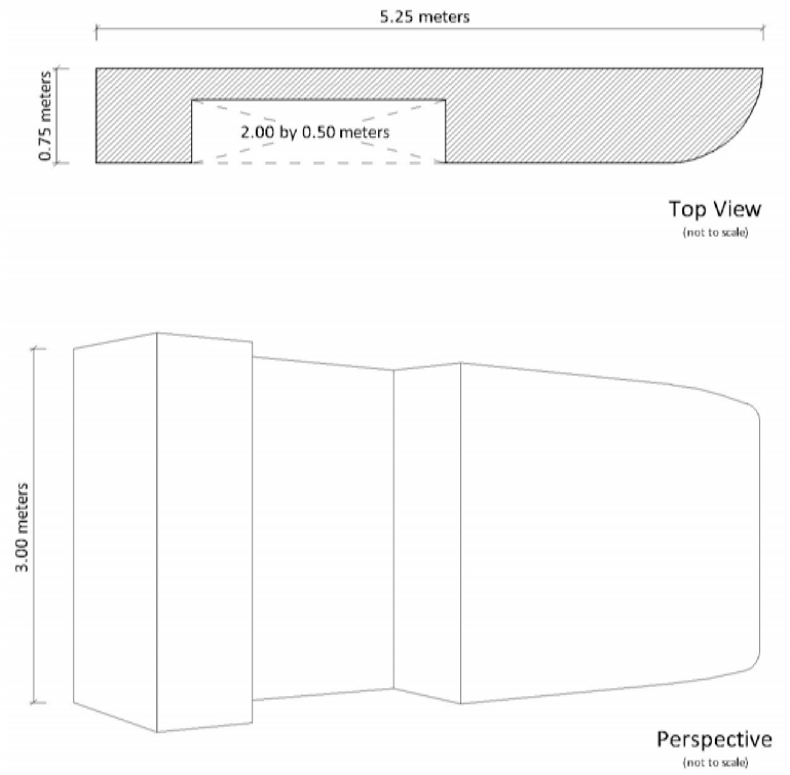



**Figure 2.** Top and perspective views of the given geometry.

The project was announced on the 21st of February 2014. Frameworks of interaction, communication, contribution and submission were explained in detail and a Q&A session followed. In addition there was a 4 day orientation period for the treatment group. The course of the project took 6 weeks and ended on the 10th of April 2014. The interactions of the treatment group was limited to strictly on Facebook whereas the control group attended a weekly 4 hour traditional design studio session.

There were two important limitations to this study. Due to initial connection problems some students failed to maintain a continuous access to Facebook group. These students have expressed that connection problems affected their performance and adaptation process. Second important limitation was the scale of the given design problem. Compared to 4th year design studio courses in other programs such as architecture and interior architecture, not only the overall scale of the project was small but also submission requirements were limited. This was to ensure that the online communication process while enabling the collection of more precise and relevant data.

**Analysis and Discussion**

The treatment and the control groups were compared using two distinct data sets. These data sets were: student activities throughout the course of the project and evaluation data of student submissions. The data didn't conform to normal distribution. Therefore 6 different non-parametric tests that are alternatives for the independent t-test were utilized (Wald-Wolfowitz Runs Test, Median Test, Mann-Whitney U Test, Kolmogorov-Smirnov Test, Kruskal-Wallis Test, and Jonckheere-Terpstra Test for Ordered Alternatives). Although all tests were pointing to same results, Mann-Whitney U test was employed as this test demonstrated a consistency with results.

Even though homogeneity tests couldn't be performed on the data, the student activity graphs shows parallels between both groups (Table 1, Figure 3). The density of activities for the treatment group suggests that the orientation and adaptation period was fairly short, implying that the treatment group has understood the dynamics of interaction and started acting accordingly in a short time. This might be due to the familiarity with Facebook interface. The graph shows that the treatment group has generated considerably more activity, however it should be noted that critiques sessions were longer for the control group. Considering this was the only course conducted over social media for the treatment group, it should also be noted that students' focus might shift and the amount of activity might decrease in case of additional courses conducted on Facebook.

When the contours created by activities from both groups are examined it can be seen that the density of activities increase and decrease similarly. For instance during the first week there was an increase in activities for both groups but during the second week activities diminished for both groups. Again there was a rise of activities on the third week again diminishing during the 4th week, which is the midterm week for the semester. This indicate that using the online continuous interaction platform won't change studying habits and dispersion of workload, contrary to expectation.

Student submissions were graded in four categories conceptualization/form, functionality/ergonomics, originality/creativity, and visuals/presentation over a 10 point scale. The results of analysis of each category and overall success are provided below (Table 2). In terms of general success there is a significant difference between the treatment and the control group suggesting that the perks provided by the continuous interaction platform such as ease of communication, high exposure to peer progress, ease of archiving and backtracking and freedom from time/space limitations. Also when comparisons between each category reveals that, there was a meaningful difference between the treatment and the control group in conceptualization/form and functionality/ergonomics categories, however there was no significant difference between both groups in originality/creativity and visuals/presentation categories.

**Table 1.** Student activity table.

| | Treatment | | | | | Control | |
|---|---|---|---|---|---|---|---|
| | Like | Comment | Share | Critiques | Attendance | | |
| 23 -Feb - Sunday | 114 | 5 | 19 | | | | |
| 24 - Feb - Monday | 194 | 21 | 29 | | | | |
| 25 - Feb - Tuesday | 237 | 33 | 27 | | | | |
| 26 - Feb - Wed | 274 | 32 | 20 | | | | |
| 27 -Feb - 2014 | 124 | 22 | 13 | 5 | 30 | | |
| 28 - Feb - 2014 | 2 | 4 | 0 | 2 | 0 | | |
| 01 - March - 2014 | 14 | 1 | 1 | 0 | 0 | | |
| 02 - March- 2014 | 32 | 3 | 2 | 0 | 0 | | |
| 03 - March - 2014 | 17 | 1 | 1 | 1 | 0 | | |
| 04 - March - 2014 | 10 | 6 | 1 | 1 | 0 | | |
| 05 - March - 2014 | 110 | 24 | 6 | | 0 | | |
| 06 - March - 2014 | 320 | 55 | 16 | 18 | 22 | | |
| 07 - March - 2014 | 138 | 32 | 5 | 2 | 0 | | |
| 08 - March - 2014 | 223 | 33 | 10 | 0 | 0 | | |
| 09 - March - 2014 | 161 | 34 | 8 | 0 | 0 | | |
| 10 - March - 2014 | 95 | 15 | 2 | 0 | 0 | | |
| 11 - March - 2014 | 61 | 7 | 3 | 0 | 0 | | |
| 12 - March - 2014 | 241 | 45 | 7 | 4 | 0 | | |
| 13 - March - 2014 | 201 | 55 | 5 | 16 | 28 | | |
| 14 - March - 2014 | 99 | 29 | 5 | 1 | 0 | | |
| 15 - March - 2014 | 50 | 11 | 1 | 0 | 0 | | |
| 16 - March - 2014 | 220 | 68 | 5 | 0 | 0 | | |
| 17 - March - 2014 | 157 | 30 | 4 | 1 | 0 | | |
| 18 - March - 2014 | 43 | 9 | 2 | 0 | 0 | | |
| 19 - March - 2014 | 65 | 4 | 3 | 0 | 0 | | |
| 20 - March - 2014 | 237 | 37 | 8 | 6 | 16 | | |
| 21 - March- 2014 | 70 | 18 | 2 | 0 | 0 | | |
| 22 - March - 2014 | 26 | 3 | 1 | 0 | 0 | | |
| 23 - March - 2014 | 0 | 0 | 0 | 0 | 0 | | |
| 24 - March - 2014 | 49 | 4 | 2 | 0 | 0 | | |
| 25 - March - 2014 | 88 | 15 | 4 | 0 | 0 | | |
| 26 - March - 2014 | 1 | 2 | 0 | 0 | 0 | | |
| 27 - March - 2014 | 53 | 7 | 2 | 4 | 18 | | |
| 28 - March - 2014 | 0 | 0 | 0 | 1 | 0 | | |
| 29 - March - 2014 | 88 | 4 | 3 | 0 | 0 | | |
| 30 - March - 2014 | 24 | 0 | 1 | 0 | 0 | | |
| 31 - March - 2014 | 10 | 5 | 0 | 0 | 0 | | |
| 01 - April - 2014 | 0 | 0 | 0 | 0 | 0 | | |
| 02 - April - 2014 | 22 | 0 | 1 | 0 | 0 | | |
| 03 - April - 2014 | 105 | 18 | 6 | 24 | 28 | | |
| 04 - April - 2014 | 127 | 27 | 3 | 2 | 0 | | |
| 05 - April - 2014 | 139 | 23 | 5 | 0 | 0 | | |
| 06 - April - 2014 | 181 | 43 | 6 | 0 | 0 | | |
| 07 - April - 2014 | 104 | 19 | 5 | 1 | 0 | | |
| 08 - April - 2014 | 279 | 33 | 11 | 0 | 0 | | |
| 09 - April - 2014 | 572 | 67 | 28 | 3 | 0 | | |
| 10 - April - 2014 | 249 | 52 | 16 | 0 | 0 | | |
| | **5626** | **956** | **299** | **92** | **142** | | |

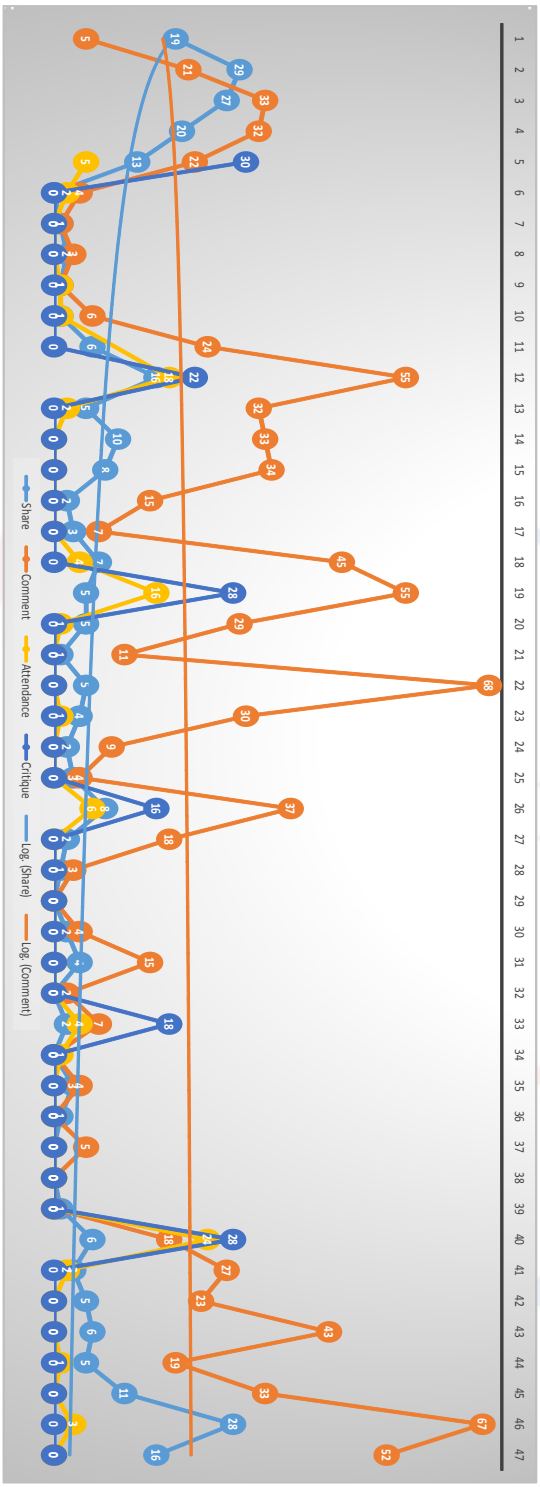



**Figure 3.** Student activity graph.

**Table 2.** Analysis of results for each category.

| Analysis Category | Sig. Value |
|---|---|
| **Overall Success** | **0.000** |
| **Conceptualization/Form** | **0.003** |
| **Functionality/Ergonomics** | **0.013** |
| Originality/Creativity | 0.072 |
| Visuals/Presentation | 0.069 |

The conceptualization/form category signifies the ability of student to analyze, understand and implement visual features of the selected periodical, reflecting visual cues while creating a coherent whole. The significant difference between the treatment and the control group might be a result of the access available to conceptualization processes of peers and the judgment free environment for analysis. The functionality/ergonomics category signifies the ability of student to provide functionality that is also ergonomic for given requirements. There is a significant difference between the treatment and control group in this second category. This might be due to the archiving and backtracking functionality that is built into the proposed continuous interaction platform. Also it should be considered that during the study session students had complete access to search engines unlike the traditional design studio environment. Treatment group could refer to professional examples and peers' solutions whereas the control group had limited access to such data.

On the other hand, there was no significant difference between the treatment and contol group for the originality/creativity category which focused on the originality of conceptual, formal and functional solutions proposed by the student. The lack of significant difference might imply that contrary to expectations, that is increased peer plagiarism would be encountered on the continuous interaction platform, the design solutions provided by students were sufficiently varied and original across the board. The visuals/presentation category was an assessment of the skills for preparing visuals and presentations. The lack of significant difference between the treatment and the control group for this category might be a result of the fact that the skills that are required for this category weren't addressed throughout this course. Also shared submission visuals failed to stimulate inspiration and a significant difference between groups.

## Conclusions

Findings of this study suggests that implementing a social media based continuous interaction platform for design studio courses had created a positive impact on students' overall performance. It should be noted that the scale and the scope of the given design problem and submission requirements were modest, therefore current functionality of Facebook sufficiently supported the proposed continuous interaction platform. However, a larger scale and scope might require new tools and a new interaction framework. An alternative to the proposed continuous interaction platform could be a hybrid process, involving traditional sessions complemented with online

sessions for which there are several successful implementations, but it should also be considered that a hybrid approach might also cause the student to focus on a single platform while not benefiting from the other (English and Duncan-Howell, 2008; Longbottom, 2008; Tate and Osborne, 2013). Another improvement for the continuous interaction platform might be an open group approach. Anybody outside of the class roster can access and contribute to the discussion. However it should also be noted that open groups might cause increased levels of social discomfort, indifference towards interaction and pedagogical lurking (Kreijns et. al., 2007). Another point of concern for a continuous interaction platforms is the possible extensity of social media use in learning in the future. Although learning process is manageable with a single course, students might struggle to maintain focus with several courses on social media platforms.

In conclusion the findings of this study complements the growing body of research on the use of social media in learning environments. Although existing research on the subject have provided a solid foundation for this particular research, a complete implementation of social media learning, especially in the field of design education would require further research focusing on learning scenarios of differing context and scales. As the advantages and disadvantages of social media implementation in learning becomes clearer, it can only be expected that social media would become an intricate part of education with added features and tools specific for the task.

## References


Anklam, P. (2009). *Ten Years of Network*. The Learning Organization, 16(6), 415-426.

Bayraktutan, G. (2013). Sosyal Medyada Etkileşim in (Öztürk, M.C. Editor). *Dijital İletişim ve Yeni Medya* (s. 100-119). Eskişehir/Turkey: Anadolu University Press.

Boyd, D.M. and Ellison, N.B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13(1), s. 1-?

Boyd. D. (2012). Participating in the Always-On Lifestyle *in The Social Media Reader* (s. 71-76)(Ed: M. Mandiberg). New York: New York University Press.

Brusasco, P.L., Caneparo, L., Carrara, G., Fioravanti, A., Novembri, G. ve Zorgno A.M. (2000). Computer Supported Design Studio. *Automation in Construction*, 9, s. 393-408.

Cochrane, T., and Bateman, R. (2010). Smartphones Give You Wings: Pedagogical affordance of mobile web 2.0. *Australasian Journal of Educational Technology*, 26(1), 1-14.

Conole, G. and Culver, J. (2010). The Design of Cloudworks: Applying Social Networking Practice to Foster the Exhchange of Learning and Teaching Ideas and Designs. Computers & Education. 54, s. 679-692.

Dickey M.D. (2007). Teaching in 3D: Pedagogical affordances and constraints of 3D virtual worlds for synchronous distance education, *Distance Education*, 24, 105-121.

Given, L. M. 2008. The Sage Encyclopedia of Qualitative Research Methods. Thousand Oaks, CA: Sage.

Goodchild, M.F. (2007). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research* Vol 2, s. 24-32.

English, R., and Duncan-Howell, J. (2008). Facebook Goes to College: Using social networking tools to support students undertaking teaching practicum. *Merlot Journal of Online Learning and Teaching*, 4, p. 596- 601. Retrieved 23/03/11

Hoyt B. (2000). Techniques to manage participation and contribution of team members in virtual teams. *WebNet Journal*, 2, 16-20.

Jonassen D.H. (1999). Designing constructivist learning environments, (Reigeluth C.M. editor), *Instructional design theories and models*, Mahwah, NJ, Erlbaum Associates, 215-239.

Jones, S. (1997). *Virtual Culture*. London: Sage.

Kaplan, A.M. and Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons* 53, s. 59-68.

Komito, L. and Bates, J. (2009). Virtually Local: Social Media and Community Among Polish Nationals in Dublin. *Aslib Proceedings: New Information Perspectives*, 61(3), 232-244.

Kreijns, K., Kirschner, P. A., Wim, J., and van Buuren, H. (2007). Measuring perceived sociability of computer-supported collaborative learning environments. *Computers & Education*, 49(2), 176–192.

Kurt S. (2009). An Analytic Study on the Traditional Studio Environments and the Use of Constructivist Studio in the Architectural Design Education. *Procedia Social and Behavioral Sciences*, 11, p. 401-408.

Longbottom, C. (2008). Pilot study for integrated blended learning in a first year studio design program, STP. Retrieved 20/04/2011

Martin J. R., Ed. (2001). *Language, register and genre: Analysing English in a global context.* London, Routledge.

Mazmani S.G. and Usluel, Y.K. (2010). Modeling Educational Usage of Facebook. *Computers & Education* 54, s. 444-453.

McLoughlin, C., and Lee, M. J. W. (2007). Social software and participatory learning: pedagogical choices with technology affordances in theWeb 2.0 era. Paper presented at the Ascilite, Singapore.

McLaughlin, Patricia and Mills, Anthony (2008), Where shall the future student learn? Student expectations of university facilities for teaching and learning, in *Proceedings of the 17th Annual Teaching and Learning Forum*, [Curtin University], [Perth, W.A.], pp. 1-9.

O'Reilly, T., (2005), O'Reilly — What Is Web 2.0, retrieved in 15/4/2011, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.

O'Reilly, T., (2007), What Is Web 2.0: Design Patterns and A Business Models for the Next Generation of Software, retrieved in 15/02/2011 http://mpra.ub.unimuenchen.de/4578/1/MPRA_paper_457

Özata, F.Z. (2013). Sosyal Medya Platformları. in Öztürk, M.C. (Ed.). Dijital İletişim ve Yeni Medya (s. 76-98). Eskişehir/Turkey: Anadolu University Publishing

Özünlü, S. (2012). *Social Media As a Consumer Socialization Agent for Adolescents and Adolescents' Purchase Intention Between the Ages of 11-15*. Unpublished Master's Thesis. İzmir: Yeditepe University.

Rheingold, H. (2000) *The virtual community*. Cambridge, MA: MIT Press.

Robins, K. and Webster, F. (1999). *Times of the Technoculture*. London: Routledge.

Robinson, J. E.(2007). A Study of Social Media Marketing In North Carolina Special Libraries. Unpublished Master's Thesis, University of North Carolina at Chapel Hill, U.S.A.

Salkind, N. J. 2006. *Exploring Research.* New Jersey: Pearson Education International.

Schön, D.A. (1987). *Educating the Reflective Practitioner*. San Francisco: Jossey-Bass.

Singh, B. (2010), *Difference Between Web 1.0, Web 2.0, & Web 3.0 - With Examples*, Retrieved 30/5/14 http://ezinearticles.com/?Difference-Between-Web-1.0,-Web-2.0,-and-Web-3.0---With-Examples&id=3683790.

Stark, E. (2006) "Free Culture and the Internet: A New Semiotic Democracy" openDemocracy.net, Retrieved 20/6/2014 http://opendemocracy.net/arts-commons/semiotic_3662.js

Tate, D. and Osborne, L. (2013). *Facebook: A 24 Hour Studio Environment for Contemporary Architectural Education*. [Working Paper] (Unpublished). retrieved 12/01/2014 http://eprints.qut.edu.au/59549/

Tu, C. H., and Isaacs, M. (2002). An examination of social presence to increase interaction in online classes. *American Journal of Distance Education*, 16(3), 131–150.

Uden, L. (2007). Activity theory for designing mobile learning. *International Journal of MobileLearning and Organisation,* 1(1), 81-102.

Urista, M. A., Q. Dong and K. D. Day. (2009). Explaining Why Young Adults Use MySpace and Facebook Through Uses and Gratifications Theory. *Human Communication, A Publication of the Pacific and Asian Communication Association*. Vol. 12, No. 2, pp.215 - 229.

Utaberta, N., Hassanpour, B., Che Ani, A.I. and Surat, M. (2011). Reconstructing the Idea of Critique Session in Architecture Studio. *Procedia Social and Behavioral Sciences* 18, s. 94-102.

iafor

# *Applying the ICA-Based Approach to Detect Faults in Processes*

Mu-Chen Chen, National Chiao Tung University, Taiwan, R.O.C.
Chun-Chin Hsu, Chaoyang University of Technology, Taiwan, R.O.C.
Bharat Malhotra, Indian Institute of Technology, India
Manoj Kumar Tiwari, Indian Institute of Technology, India

The Asian Conference on Technology, Information & Society
Official Conference Proceedings 2014

**Abstract**
To save time, cost and labor, there are many studies that have been conducted about the detection of faults in industrial processes. Most of the previous studies used only Independent Component Analysis (ICA) or Principal Component Analysis (PCA) for detection, but they cannot form close enough boundaries to reject outliers. This paper proposes an ICA-based approach to detect outliers in a process by forming close boundaries. The basic idea of the proposed method is to apply ICA to convert original data into independent components, and then apply Durbin Watson (DW) criterion to select important independent components. Hereafter, Support Vector Data Description (SVDD) has been applied for outlier detection by forming much tighter boundaries. The efficiency of proposed ICA-based approach is investigated via a simulated multivariate process example. To demonstrate the identification capability of the proposed ICA-based approach, the traditional Hotelling's $T^2$ chart is constructed for the simulated data set.

Keywords: Fault detection, Independent Component Analysis, Principal Component Analysis, Hotelling's $T^2$ chart

## Introduction

In Statistical Process Control (SPC), Hotelling's $T^2$ chart adopts the original variable to calculate Mahalanobis distance. A simple type of Multivariate Statistical Process Control (MSPC) is a multivariate chart which is an extended from univariate SPC methods like Hotelling's $T^2$ chart (Montgomery, 2001). This chart is unable to detect small changes taking place during the cycle. To remove this drawback, MEWMA and MCUSUM are developed to detect small process changes (Virojana *et al*., 2003). However, the fault detectability decreases as the number of variables increases. Principal Component Analysis (PCA) has been widely used for detecting faults because of its compatibility with many of the methods available in MSPC (Bakshi, 1998) and its ability to reduce attribute dimensions. However, PCA lacks in assuming state of variable, it assumes that latent variable follows Gaussian distribution, while in chemical processes variables hardly remains at stable state because of the uncontrolled disturbances.

Dynamic PCA (DPCA) has been proposed which uses an amplified matrix with time lagging variables (Ku *et al*., 1995). Serial correlation of data has been considered in this method. Along with other developed methods, DPCA has been widely used in many fields such as sensor fault detection (Luo *et al*., 1999 and Rato *et al*., 2013), multi-scale fault detection, multi-scale analysis, simultaneously monitoring, diagnosis in the wastewater treatment process (Yoo *et al*., 2002) and diagnosing an automatic controlled process (Tsung, 2000). Kernal PCA has been developed for nonlinear nonstationary process monitoring (Khediri *et al*., 2011).

Techniques based on Independent Component Analysis (ICA) have been developed (Kano *et al*., 2003 and Lee *et al*., 2003) recently. ICA is used for decomposition of data into linear combination of components independent from each other. Then, ICA monitoring on lagged variables, Dynamic ICA has also been used to detect faults. Chiang *et al*. (2004) has used Support Vector Machine (SVM) and Fisher discriminant analysis (FDA) for fault diagnosis. Guo *et al*. (2014) proposed envelope based dimension reduction for ICA in fault diagnosis.

ICA has been used extensively for fault detection since it gives more sophisticated results than PCA, and the ICA's power in fault detection has been ascertained. However, the improvement of ICA based fault detection approach is still in need because of the following limitations of ICA (Hyvärinen and Oja, 1997; Hyvärinen, 1999; Hsu *et al*., 2010): 1) the incapability of identifying important independent components (ICs), 2) the lack of understanding the influence of original variable on given IC, and 3) the loose boundary for enveloping all data points.

Understanding the abovementioned limitations of ICA, this paper aims at developing an efficient approach of fault detection and diagnosis for non-Gaussian processes. In the proposed approach, important ICs are selected with the help of Durbin Watson (DW) criterion (Dublin and Watson, 1950). Original variables are retraced to obtain a view that which original variable is influencing the identified important independent components. Further, Support Vector Data Description (SVDD)(Tax and Duin, 2004) has been applied for the purpose of tightening of the boundary surrounding the data points.

## The Proposed Fault Detection and Diagnosis Approach

Firstly, the basics of ICA, DW and SVDD are presented. Independent component analysis (ICA) is an approach which supposes statistical independence of non-Gaussian source signals for dividing a multivariate signal into smaller components. Readers are referred to (Lee *et al*., 2003) for the further details of ICA. ICA can be of many types such as Jade (Rutledge and Bouveresse, 2013), FastICA (Hyvärinen, 1999), etc. In this paper, we use FastICA.

Hyvärinen (1999) invented FastICA which is an efficient algorithm for ICA. FastICA has many advantages over normal ICA model which are discussed later in this paper. Centering and whitening of the input vector data x has to be done before applying FastICA algorithm (Hyvärinen, 1999). The FastICA is finds maximum of the non-Gaussianity (Hyvärinen, 1999; Hyvärinen and Oja, 1997) and it is based on fixed point iteration scheme. By approximative Newton iteration FastICA can be derived.

Recently, Durbin Watson (DW) criterion has been widely used in many models, and recently found its application in ICA (Ammari *et al*., 2011). Initially this methodology was proposed for measuring signal/noise ratio (Dublin and Watson, 1950). In absence of any king of noise, the DW value tends toward 0, and if signal consist of only noise, it will incline to 2. This criterion has been used for validation of the multivariate models (Rutledge *et al*., 2002; Gourvénec *et al*., 2002 and Gomez-Carracedo *et al*., 2007). Readers are referred to Dublin and Watson (1950) for the further details of DW.

In Support Vector Data Description (SVDD), it is assumed that the data is enclosed by a hyper sphere with minimum volume (Tax and Duin, 2004). We try to minimize probability of accepting outliers by minimizing the volume of the enclosed space. SVDD is able to make tighter boundaries around the data points. Readers are referred to (Tax and Duin, 2004) for the further details.

The work of SVDD is to map the target data nonlinearly into a higher dimensional feature space and construct a separating hyperplane with maximum margin there. It is probable to find the dividing hyperplane without explicitly carrying out the map into the feature space by using kernel function. The dot product $x_i x_j$ can also be avoided by implementing kernel function. Kernel function is any kind of a function that follows Mercer's Theorem (Sch¨olkopf *et al*., 1999). The most often used kernels are polynomial kernel and radial basis function (RBF) (Sun and Tsung, 2003). In this paper, RBF has been used as the kernel.

After introducing the basics of ICA, DW and SVDD, we describe the architecture of the proposed fault detection and diagnosis approach for non-Gaussian processes. The algorithmic procedure for ICA-DW-SVDD consists of two phases, offline training and online monitoring. After data preprocessing which includes centering and whitening of data the much advanced algorithm FastICA is applied first to obtain ICs. A large number of ICs causes involvement of noise in the data, therefore the DW algorithm is applied to get important ICs. They can be selected by observing the DW color plot. For training purpose selected important ICs are used to obtain SVDD parameters which are further used while monitoring of testing data. SVDD algorithm is used enclosing data points in much tighter boundary and for detecting outliers.

After detection of outliers, original variables are retraced in order to get a clear view of the factors affecting faults, and this is done with the help of proposed retracing algorithm.

**A Simulation Example**

In this paper, the proposed approach is applied to monitor a simple multivariate process, and this simulation work is similar to (Lee *et al*., 2004 and Ku *et al*., 1995) in which there are five monitored variables in a dynamic process as follows:

$$Z(k) = \begin{bmatrix} .118 & -.191 & .287 \\ .847 & .264 & .943 \\ -.333 & .514 & -.217 \end{bmatrix} \times Z(k-1) + \begin{bmatrix} 1 & 2 \\ 3 & -4 \\ -2 & 1 \end{bmatrix} \times U(k-1)$$

$$Y(k) = V(k) + Z(k)$$

where $Y$ is the output with three variables ($y_1$, $y_2$, $y_3$). $Y$ is the normal distributed random vector with zero mean and variance of 0.1. $U$ is the input with

$$U(k) = \begin{bmatrix} .811 & -.226 \\ .477 & .415 \end{bmatrix} \times U(k-1) + \begin{bmatrix} .193 & .689 \\ -.320 & -.749 \end{bmatrix} \times W(k-1)$$

where $W$ is a random vector following uniform distribution over interval (-2, 2). Input is $U$ and output is $Y$, for process monitoring total five variables ($y_1$, $y_2$, $y_3$, $y_4$, $y_5$) are used.

A total of 1,000 observations are sampled for each simulation. The first 500 observations are used as a training data set and the remainders are used for on-line process monitoring. At observation 500, a step change of $w_1$ by 3 is introduced. This means that the first 500 training observations are not contaminated by outliers, and the remaining 500 data are faulty. The FastICA algorithm is applied to the generated data, and 5 ICs are obtained as a result of this. Results of DW criterion represents that 4 ICs are important, while 1 IC contains mainly noise. The 4 important ICs have been divided into target and outlier data sets. First 500 data are selected for testing, and the rest for outlier detection. In the training data, no outlier is selected. And results are obtained for number of normal data in test data as shown in Table 1. Actually, there is no normal data in test data.

Table 1: The results of simulation example.

| Value of sigma | Detection rate (%) |
|---|---|
| 3 | 99 |
| 4 | 100 |
| 5 | 100 |
| 6 | 98 |
| 7 | 96 |
| 8 | 96 |
| 9 | 94 |
| 10 | 92 |

There, for sigma equal to 4 and 5, error is equal to 0, i.e. every abnormal data has been verified as outlier data. We choose the value of sigma equals to 4 for further calculation. The hyper spherical boundary rejects 500 outliers, and makes tighter boundary enclosing only 500 normal data. Figure 5 illustrates the graph of distances of generated data from the centroid of the data points. Width parameter of RBF kernel function's is found out using sensitivity analysis and using sigma value equal to 4 in SVDD, it itself sets threshold according to the number of abnormal data we consider for training. It has been seen in Figure 1 that threshold has been set around 0.71 by SVDD to enclose all normal data and hence rejects all other points above threshold. Those points above the threshold should be considered as faulty.

## Conclusion

In this paper, we proposed an approach based on ICA, DW and SVDD for the purpose of fault detection and diagnosis in non-Gaussian processes. The fault detection problem is converted to a one-class data description problem. The proposed monitoring method uses FastICA to get ICs. The DW algorithm is used to get important ICs, which further used in SVDD algorithm for enclosing data points in much tighter boundary and for detecting outliers. The model based on SVDD expresses fault-free data distribution using a hyper sphere with a close-fitting boundary. The proposed approach is validated on a simulation example.

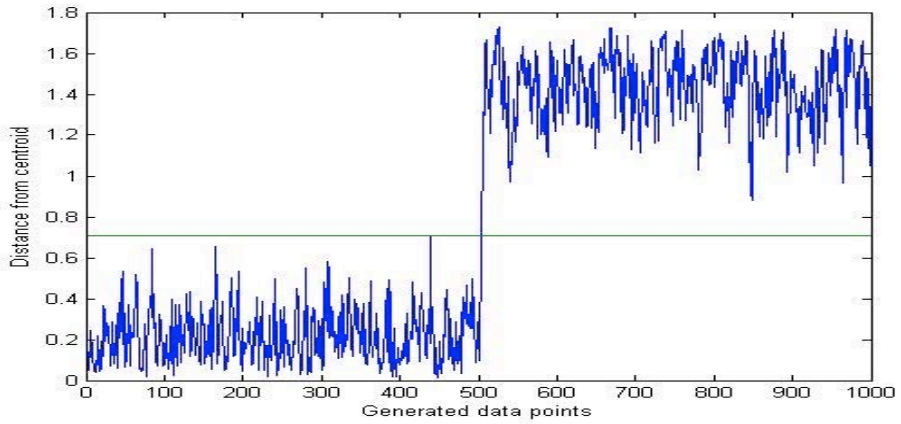



Figure 1: The distances of generated data from the centroid of data points.

## Acknowledgements

This work is partially supported by National Science Council, Taiwan, R.O.C under grant NSC 100-2410-H-009-013-MY3.

## References

Ammari, F., Cordella, C.B.Y., Boughanmi, N. and Rutledge D.N., 2011. Independent Component Analysis applied to 3D-front face fluorescence of spectra of edible oils to study the antioxidant effect of Nigella sativa L. extract on the thermal stability of heated oils, *Chemometrics and Intelligent Laboratory Systems*.

Bakshi, B.R., 1998. Multiscale PCA with application to multivariate statistical process monitoring, *American Institute of Chemical Engineering Journal* 44 (7), 1596–1610.

Chiang, L. H., Kotanchek, M. E. and Kordon A. K., 2004. Fault diagnosis based on Fisher discriminant analysis and support vector machines, *Computers & Chemical Engineering*, 28(8), 1389-1401.

Hsu, C.C., Chen, M.C and Chen, L.S., 2010. A novel process monitoring approach with dynamic independent component analysis, *Control Engineering Practice* 18(3), 242–253.

Durbin, J. and Watson, G.S., 1950. Testing for serial correlation in least squares regression, *Biometrika*, 37, 409–428.

Gomez-Carracedo, M.P., Andrade, J.M., Rutledge, D.N. and Faber, N.M., 2007. Selecting the optimal number of partial least squares components for the calibration of attenuated total reflectance mid-infrared spectra of undersigned kerosene samples, *Analytica Chimica Acta*, 585, 253-265.

Gourvénec, S., Massart, D.L. and Rutledge, D.N., 2002. Determination of the number of components during mixture analysis using the Durbin–Watson criterion in the orthogonal projection approach and in the simple-to-used interactive self modeling mixture analysis approach, *Chemometrics and Intelligent Laboratory Systems*, 61, 51-61.

Guo, Y., Na, J., Li, B. and Fung, R.F., 2014. Envelope extraction based dimension reduction for independent component analysis in fault diagnosis of rolling element bearing, *Journal of Sound and Vibration*, 333(13), 2983-2994.

Hyvärinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis, *IEEE Transactions on Neural Networks*, *10* (3), 626-634.

Hyvärinen, A. and Oja, E., 1997. A fast fixed-point algorithm for independent component analysis, *Neural Computation*, 9(7), 1483-1492.

Kano, M., Tanaka, S., Hasebe, S., Hashimoto, I. and Ohno, H., 2003. Monitoring independent components for fault detection, *American Institute of Chemical Engineering Journal*, 49 (4), 969-976.

Khediri, I.B., Limam, M. and Weihs, C., 2011. *Variable window adaptive Kernel Principal Component Analysis for nonlinear nonstationary process monitoring*, Computer and Industrial Engineering, 61(3), 437-446.

Ku, W., Storer, R.H. and Georgakis, C., 1995. Disturbance detection and isolation by dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, 30, 179-196.

Lee, J.M., Qin, S.J. and Lee, I.B., 2007. Fault detection of non-linear process using kernel independent component analysis, *The Canadian Journal of Chemical Engineering*, 85, 526-536.

Lee, J.M., Yoo, C.K. and Lee, I.B., 2003a. Statistical process monitoring with independent component analysis, *Journal of Process Control*, 14 (5), 467-485.

Lee, J.M., Yoo, C.K. and Lee, I.B., 2003b. Statistical process monitoring with multivariate exponentially weighted moving average and independent component analysis, *Journal of Chemical Engineering of Japan*, 36 (5), 563-577.

Lee, J.M., Yoo, C.K. and Lee, I.B., 2004b. Statistical monitoring of dynamic processes based on dynamic independent component analysis, *Chemical Engineering Science*, 59, 2995-3006.

Lee, J.M., Yoo,C.K. and Lee,I.B. 2004a. Statistical process monitoring with independent component analysis, *Journal of Process Control*, 14, 467-485.

Luo, R., Misra, M. and Himmelblau, D.M., 1999. *Sensor fault detection via multiscale analysis and dynamic PCA*, *Industrial and Engineering Chemistry Research*, 38, 1489-1495.

Montgomery, D. C., 2001, *Introduction to Statistical Quality Control*, 4th ed., New York: Wiley.

Rato, T.J. and Reis, M.S., 2013. Fault detection in the Tennessee Eastman benchmark process using dynamic principal components analysis based on decorrelated residuals (DPCA-DR), *Chemomatrics and Intelligent Laboratory Systems*, 125,101-108.

Rutledge, D.N. and Barros, A.S., 2002. Durbin–Watson statistic as a morphological estimator of information content, *Analytica Chimica Acta*, 454, 277-295.

Rutledge, D.N. and Bouveresse, D.J.R., 2013. *Independent Component Analysis with Jade Algorithm*, 50, 22-32.

Sch¨olkopf, B., Burges, C.J.C. and Smola, A.J. 1999. *Advances in Kernel Methods.* The MIT Press.

Sun, R. and Tsung, F., 2003. A kernel-distance-based multivariate control chart using support vector methods, *International Journal of Production Research*, 41(13), 2975-2989.

Tax, D.M.J and Duin, R.P.W., 2004. Support vector data description, *Machine Learning*, 54, 45-66.

Tsung, F., 2000. Statistical monitoring and diagnosis of automatic controlled processes using dynamic PCA, *International Journal of Production Research*, 38 (3), 625-637.

Yoo, C.K., Choi, S.W. and Lee, I.-B., 2002. Dynamic monitoring method for multiscale fault detection and diagnosis in MSPC, *Industrial and Engineering Chemistry Research*, 41, 4303-4317.

**Contact email:** ittchen@mail.nctu.edu.tw

iafor

The International Academic Forum (IAFOR)
Sakae 1-16-26-201
Naka Ward, Nagoya, Aichi
Japan 460-0008
ww.iafor.org