IT&T

The IT&T
13th International Conference on Information Technology and Telecommunication
2014

Towards the Internet of Things: Issues, Challenges and Approaches

8 - 9 May 2014

Markus Hofmann
Gabriel-Miro Muntean
(Eds)

ISSN 1649-1246

IT&T 2014 General Chairs' Letter

As Chairpersons of the 13th Information Technology and Telecommunications Conference (IT&T 2014), we have great pleasure in welcoming you to this year's conference which will be hosted by the School of Electronic Engineering at Dublin City University on the 8th and 9th of May 2014.

This year's conference theme is "Towards the Internet of Things: Issues, Challenges and Approaches". Originally proposed to support the mobile computing requirements of the "dot com" era, enhanced mobile telecommunication standards were generational in nature. Over a decade later, mobile devices such as the iPhone, iPad and Android smart phones mean such networks are receiving the level of utilisation originally anticipated. The interactivity of such devices consumes and generates significant amounts of disparate data: multimedia data, performance data, sensory data, etc. The effective creation, transmission and analysis of such data provides real challenges for the research and development community.

This year's conference includes 17 research papers and several posters from research from Irish Institutes of Technology and Universities. The papers and posters will be presented in six sessions. The key-note speeches will be delivered by Dr. Maria Moloney from Escher Group and Thorsten Dahm from Google. Additionally we will have a Doctoral Consortium.

Many people have helped with the preparation of this year's conference. A big thank you for the very enthusiastic support of the local organising committee team, under the excellent lead of Dr. Zhenhui Yuan and Dr. Olga Ormond, which has included administrators, technicians and postgraduate researchers from School of Electronic Engineering and the Rince Institute at Dublin City University. Last, but not least, thank you also to the session chairs and doctoral consortium panellists for helping run the conference programme.

We herewith welcome you to Dublin City University and to the IT&T 2014 conference.

Dr Gabriel-Miro Muntean and Dr. Markus Hofmann,
General Chairs IT&T 2014, Dublin, Ireland

Technical Programme Committee Chairs' Letter

Dear Colleagues,

As Technical Programme Chairs, we would like to welcome you to the Thirteenth Information Technology and Telecommunications Conference (IT&T 2014) hosted by the Dublin City University, Ireland.

IT&T is an annual conference which not only publishes research in the areas of information technology and telecommunications, but also brings together researchers, developers and practitioners from the academic and industrial environments, enabling research interaction and collaboration.

The focus of the Thirteenth IT&T is "Towards the Internet of Things: Issues, Challenges and Approaches". There is little doubt the Internet as we use it today more complex, versatile and integrated than ever before and changing rapidly. It also apparent that we are in the midst of a "Data Deluge" as predicted by the Economist in 2010 and IEEE in 2011 and that the current focus of many researchers is aimed at advanced analytics. The focus of this conference is how we manage and deal with this twin revolution in technology and how we can identify, prioritize and help solve the real challenges that are generated.

We welcomed research papers with topics in wireless and mobile networks, sensor networks and embedded systems, energy efficient computing and communications, ubiquitous and distributed computing, cyber physical systems, security in information and telecommunication systems, web technologies for a smarter planet, digital signal processing, adaptive computing, management of ICT systems, cloud computing and services, applications of artificial intelligence and machine learning, ICT for health, transport, traffic, water, and energy, open source development, data, text and web content mining, and last but not least, ICT for education.

All submitted papers were peer-reviewed by the Technical Programme Committee members and on behalf of the organizing committee we express our sincere gratitude to all of them for their help in the reviewing process. The outcome of the review process produced seventeen papers that were accepted and these will be presented during the seven technical sessions spanning the two days of the conference. This year's conference will also display a number of posters and a doctoral consortium.

We hope you will have a very interesting and enjoyable conference.

Prof. John Murphy, University College Dublin, Ireland

Dr. Enda Fallon, Athlone Institute of Technology

IT&T 2014 Chairs and Committees

General Chairs

Gabriel-Miro Muntean, Dublin City University
Markus Hofmann, Institute of Technology Blanchardstown

Technical Programme Committee Chair

John Murphy, University College Dublin
Enda Fallon, Athlone Institute of Technology

Patronage & Sponsor Chair

Nick Timmons, Letterkenny Institute of Technology

Organising Committee Chairs

Zhenhui Yuan, Dublin City University
Olga Ormond, Dublin City University

Doctoral Consortium Chair

Cristina Muntean, National College of Ireland

Publicity Chair

Arnold Hensman, Blanchardstown Institute of Technology

Technical Programme Committee

Paul Archbold, AIT
Brett Becker, CCT
Kevin Curran, UU
David Denieffe, CarlowIT
Enda Fallon, AIT
Sheila Fallon, AIT
Ronan Flynn, AIT
Arnold Hensman, ITB
Markus Hofmann, ITB
Gabriel Hogan, DCU
Fredrick Japhet Mtenzi, DIT
Brendan Jennings, WIT
Michael Lang, NUIG
Brian Lee, AIT
Hugh Melvin, NUIG
Jennifer McManis, DCU
Gabriel-Miro Muntean, DCU
Cristina Muntean, NCI
Phelim Murnion, GMIT
John Murphy, UCD
Niall Murray, AIT
Brian Nolan, ITB
Ciaran O'Driscoll, DIT
Olga Ormond, DCU
Declan O'Sullivan, TCD
Dirk Pesch, CIT
Nick Timmons, LYIT
Sven Van De Meer, Ericsson
Xiaojun Wang, DCU
Zhenhui Yuan, DCU

Table of Contents

Session 3: Technical Poster Session

Session 4: Research in Progress Pitch & Doctoral Symposium

Session 5: ICT in Education

Session 6: Green Computing

Session 7: Wireless and Mobile Networks

Session 1

Knowledge Discovery

Text Mining on Software Patents

Tristan O'Gorman [1], Markus Hofmann [2]

[1] Institute of Technology Blanchardstown, Dublin, Ireland
tristanogorman@gmail.com

[2] Institute of Technology Blanchardstown, Dublin, Ireland
markus.hofmann@itb.ie

Abstract

Patents are a form of intellectual property and can be assigned to various categories according to their technological features. For software patent analysts, identifying software patents can be a time consuming process as no specific classes exist for software patents. Non-software and software patent abstracts were mined through web crawling and information extraction from the U.S. Patent Office to create a binary classification learner that can categorize unseen software patents. Various text preprocessing and data visualization techniques were employed, and learners such as Naive Bayes, SVM, and various ensemble methods were experimented with. During training, the accuracy of the optimal learner model was high. However, when deployed on unseen data the results were mixed. It is felt that a more comprehensive training data set would lead to a more robust classifier and the feasibility of a real world implementation is discussed.

Keywords: patents, text mining

1 Introduction

Patents permit the owner to stop others from using, selling or importing the invention; and they can cover objects such as devices or articles, compositions of matter, methods or processes, new applications for existing devices or materials, or a products made by new processes (Poltorak and Lerner, 2002). Patents are classified according to the technical features of the invention and the World Intellectual Property Organization has developed a classification system, the International Patent Classification (IPC) system (World Intellectual Property Organization, 2013). The IPC system is accepted by most international patent bodies, and is often used in conjunction with a country or region-specific classification system. The IPC classification model is a hierarchial model where patents are assigned to one or more sections, classes, subclasses, groups, or subgroups as determined by their technical features. Within the IPC, there are 8 sections, 129 classes, 638 subclasses, 7,391 groups, and 71,437 subgroups.

Software patenting has been a contentious issue for many years and the debate has centered around whether software should be patentable (Guadamuz, 2006). Perhaps because of this, the main international patenting body, the World Intellectual Property Organization, has been reluctant to create specific classification classes for software (Macher et al., 2008), and this has proved to be one of the key challenges for software patent analysts researching new and existing software inventions (Layne-Farrar, 2005; Hall and MacGarvie, 2006). In addition, patent classes are regularly created and revised so systems that rely on IPC classes to identify software patents are not reliable. Therefore, this paper ignores the IPC class-subclass categorization of patents and proposes a system based on a text classification model that is trained on historical software patent documents to predict which new patents can be classified as software. Assuming the ultimate objective is to quickly identify software patents, regardless of their

ultimate IPC classification, such a system could be of some benefit to software patent analysts: all that is required is a binary class label model that predicts whether a patent is software or non-software.

The paper is structured as follows: Section 2 discusses related work in the area of patent categorization. In Section 3, the methodology is described. Section 4 discusses the results and Section 5 concludes the paper with some thoughts on the patent categorization exercise.

2 Background and data

This paper uses a modified version of the approach that Graham and Mowery (2003) used to gather software patents: patent submissions from software companies could heuristically be assumed to be software patents. To keep the learning model as lightweight as possible, only the patent abstracts were required for analysis. To create training data for software patents, patent documents for the top 15 software companies were required. As of 2013, the top 15 software companies in terms of software revenue were Microsoft, IBM, Oracle, SAP, Ericsson, Symantec, HP, EMC, CA Technologies, Adobe, VMWare, Fujitsu, SAS, Intuit, and Siemens (PWC, 2013).

To create training data for non-software patents, it was required to create a random selection of patent from various subclasses, and omit any subclasses identified as containing any software patents by prior research (Hall and MacGarvie, 2006; Layne-Farrar, 2005; Macher et al., 2008).

3 Related work

Some initial research into text categorization techniques for patents was carried out by Larkey (1998), where the author used K-nearest neighbour and Bayesian classifiers to automatically assign patents to the correct classes and subclasses as defined by the International Patent Classification (IPC) system. This research was further developed in Larkey (1999) where a system was created to automatically classify patents to aid search. Although the system was not fully tested for accuracy, cursory analysis showed that the accuracy of the classification system was quite low and ranged from 25% to 32%. More extensive accuracy testing was promised by the author, but this appears not to have materialized.

Work concentrating on patents from the European Patent Office was carried out by Krier and Zacca (2002), where the objective was to design a classification system so that patent applications get assigned to the most appropriate assessment departments. The accuracy of the system was reasonable, ranging between 57% and 80% depending on the organizational level that the patent should be assigned to.

Concerns from the patent analysis community about the black box nature of automatic patent classification techniques were addressed by Fattori et al. (2003), who designed a patent classification system, PackMOLE, to be used in conjunction with traditional patent analysis techniques. This system enabled practitioners to quickly analyze and categorize patents while maintaining the professional integrity of results. Wu et al. (2010) combined subject matter expert screening with a modified support vector machine algorithm, hybrid genetic-based support vector machine (HGB-SVM) and achieved an accuracy of 85% on Korean and Chinese patent documents. Incorporating human analysis was also attempted by Zhang (2014), who designed an interactive patent classification algorithm which used multi-classifiers to assign the patents to sub-classes and used the learning model developed in that stage into an enhanced classifier at the class level. Kaur and Sapra (2013) ignored the IPC classification system and attempted to classify patents according to their general technological area of relevance: software, biological, business, and chemical. The authors used principal component analysis and logistic regression to develop their model and achieved an accuracy of 100% in their experiment. However, they did not provide details of the data used so the experiment is not reproducible.

Aside from automatic text categorization of patents, text mining has been used in a variety of other patent analysis areas. Potter and Hatton (2013) analyzed the U.S. patent repository to discover research trends being explored by major technology companies. Liang and Tan (2007) mined the text of patent documents to extract information to aid with the product innovative process. This information was then transformed into keywords to aid with patent searching but the authors indicated that the process could

be enhanced through clustering to return similar problems that might be solved by existing patents. Tseng et al. (2005) developed a machine-based system to derive feature terms so that patent mapping can be done more effectively. Building on this work, Tseng et al. (2007) designed a robust patent analysis methodology based on a framework of text mining techniques such as document preprocessing, indexing, clustering, and topic mapping.

Initial work on identifying patent classes that are commonly used to classify software patents was done by Graham and Mowery (2003). The authors identified patents granted by the U.S. Patent and Trademark Office to major companies and extracted the International Patent Classification (IPC) class and subclasses assigned to these patents at that time. Bessen and Hunt (2004) took a different approach to identifying software patents: they designed a boolean query to search for keywords in the patent text. As patent classes are regularly created and revised, recent work by Hall and MacGarvie (2006) combined both approaches to identify software patents.

4 Methodology

The research was conducted based on a modification of the Cross-Industry Standard Process for Data Mining (CRISP-DM) that was developed by Miner et al. (2012) specifically for text mining projects.

4.1 Data preparation

The data was prepared using various web mining techniques, detailed in the following sections.

Preparing software patent data To create software patent data, queries were run on the U.S. Patent Office database, to find patents granted in November, 2013. The resulting URLs were used as the base URLs for web crawls. The abstract text was then extracted from the results of the web crawl. The patent abstract text data for each company was then merged and written to an Excel file: software.xlsx. To avoid company-specific bias, a maximum of 50 extracts per company were extracted. In total, 527 patents were crawled and mined.

Preparing non-software data A similar process was created for non-software patents but, instead, a random selection of patent subclasses were crawled and mined, and the abstract data was written to an Excel file: nonsoftwaredata.xlsx

Final data set The final data set was created by merging the two Excel files to create the patent.xlsx file. The patent dataset contained 513 rows with the non-software class label and 518 rows with the software class label.

4.2 Exploring the final data set

To better understand the text data, in particular the nature of text data under the software class label, data visualization techniques such as word clouds (see Figure 1), word trees, and phrase net (see Figure 2) were experimented with prior to data preprocessing using the Many Eyes application, an IBM website for data visualization (IBM, 2013).

Findings from visualization exercises Data visualization proved to be a valuable exercise to help understand the fundamental differences between the content of software and non-software patents. For non-software patents there was a significant variety in the technological areas that these patents belonged to, and the tag cloud was somewhat more cluttered as there were few words that seem to be particularly predictive of non-software patents. Among the word most associated with non-software patents were *invention*, *includes*, and *present*. For software patents, words like *data*, *information*, and *system* seemed to be more frequent. *Data* was typically used with *storage* and *system* so it was expected that multi-word

Figure 1: Word cloud for software data

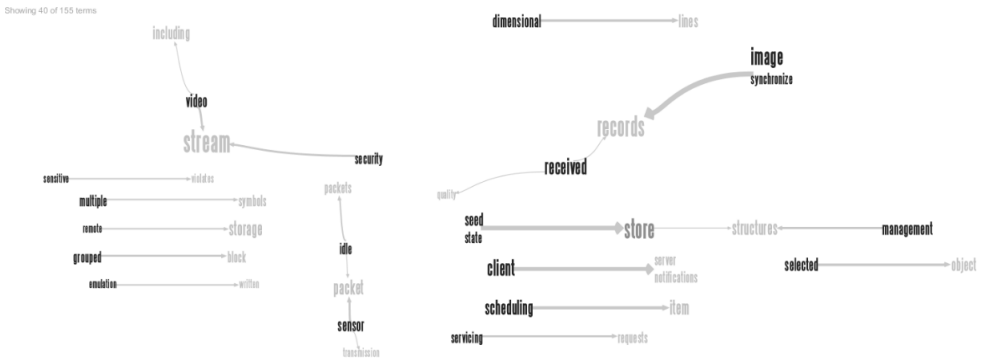


Figure 2: Phrase net for data

terms with *data* and *system* could be highly predictive. *Information* was most often used with processing but also has strong relationships with *identification* and similar terms. *System* was often used in the same context as *storage*, *file* and *computer*, and had a strong relationships with these terms. Frequent terms, such as *first*, that were common to both classes were expected to have little predictive power.

Visualizations carried out on software patent abstracts, such as phrase nets and word trees, helped the analysis of the abstract content. In particular, it seemed that certain combinations of words could be predictive of the class label. Therefore, n-gram creation was considered a worthwhile task during data preprocessing.

4.3 Preprocessing the final data set

To ascertain the efficacy of preprocessing techniques, a baseline accuracy was established using two classification algorithms: Naive Bayes and Support Vector Machines (SVM). The base algorithms were run with a Term Frequency-Inverse Document vector creation algorithm with tokenization but no pruning. The baseline accuracy was promising for both algorithms as detailed in Table 1a. The baseline accuracy was encouragingly high. As a result, it was expected that preprocessing techniques could bring only incremental improvements to the model.

Tokenization Tokenization was used to create a bag-of-words dictionary. This was done in the initial accuracy assessment and the technique created 7,965 attributes or tokens.

Transform case All words were transformed to lower case to remove any potential bias. This operator led to a slight reduction in accuracy, as can be seen in Table 1b. However, as the decrease was minimal (0.1 percentage points), it was decided to persist with case transform.

Table 1: The accuracy associated with various preprocessing techniques during the learning process

(a) Baseline Accuracy

Model	Accuracy
Naive Bayes	92.43%
SVM	93.7%

(b) Transform case

Model	Accuracy
Naive Bayes	92.34%
SVM	93.6%

(c) Filtering stopwords

Model	Accuracy
Naive Bayes	92.53%
SVM	94.18%

(d) Filtering tokens less than 2 characters

Model	Accuracy
Naive Bayes	92.53%
SVM	93.99%

(e) Bi-gram

Model	Accuracy
Naive Bayes	93.11%
SVM	93.60%

(f) Tri-gram

Model	Accuracy
Naive Bayes	93.11%
SVM	85.35%

Stopwords To remove words that have little or no predictive power, stopword filtering was employed. The U.S. patent office has a list of stopwords (U.S. Patent Office, 2010b) and this list was incorporated along with a standard English stopword list. The U.S. patent office wordlist included terms often used in patent submissions such as *embodiment*, *comprises*, *generally*, and *invention*. Stopwords filtering lead to an increased accuracy for both algorithms as detailed in Table 1c.

Filtering tokens In addition to filtering stopwords, filtering tokens of less than 2 characters was also tried. While this did not affect the accuracy of the Naive Bayes algorithm, it did lead to decrease in accuracy for SVM, as can be seen in Table 1d. Therefore, it was decided to continue filtering tokens for Naive Bayes but discontinue the practice for SVM.

N-grams To identify multi-word terms that have predictive power, bi-grams and tri-grams were generated for both algorithms. N-grams were identified during data exploration as potentially being useful for patent categorization. For Naive Bayes, both n-grams increased the accuracy. For SVM, both n-grams reduced the accuracy of the model. Therefore, it was decided to discontinue this technique for SVMs. It was decided to maintain bi-gram (see Table 1e) generation over tri-gram (see Table 1f) for Naive Bayes due to the increased computational resources required for tri-gram generation.

Stemming Stemming was attempted to improve the accuracy of the models by assuming that words with common roots will have similar meanings. Overall, stemming proved to be a productive exercise with the three stemming algorithms attempted increasing accuracy, as detailed in Table 2. For Naive Bayes, both Porters and the Snowball algorithms proved to be most accurate, while for SVMs, the Snowball algorithm was most accurate. It was decided to retain Snowball stemming for both models.

Model	Accuracy (Lovins)	Accuracy (Porters)	Accuracy (Snowball)
Naive Bayes	93.21%	93.31%	93.31%
SVM	94.09%	94.28%	94.47%

Table 2: Stemming

4.3.1 Word vectors

To find the optimal word vector representation, various word vector creation methods were attempted. In addition to TF-IDF, Term Frequency (TF), Term Occurrences (TO), and Binary Term Occurrences (BTO) were experimented with and the details can be seen in Table 3. Binary term occurrences proved most accurate for Naive Bayes, while TF-IDF was most accurate for SVM.

Model	Accuracy (TF-IDF)	Accuracy (TF)	Accuracy (TO)	Accuracy (BTO)
Naive Bayes	93.31%	93.99%	93.4%	94.08%
SVM	94.47%	93.99%	94.28%	94.46%

Table 3: Word vectors

4.3.2 Pruning

Pruning terms used in less than $x\%$ of documents was attempted to reduce the dimensionality of the dataset. Various values for x were attempted as detailed in Table 4. As can be seen, pruning served only to reduce the accuracy of the models for all values for x. Therefore, pruning was abandoned for future model development.

Model	Accuracy ($<3\%$)	Accuracy ($<2\%$)	Accuracy ($<1\%$)
Naive Bayes	91.76%	92.53%	92.34%
SVM	90.78%	92.73%	92.44%

Table 4: Pruning

4.3.3 Feature selection

In addition to pruning, various feature selection methods were attempted to reduce the dimensionality of the data set. Forward selection and backward elimination was not possible due to the lack of computing resources and a similar problem was encountered for PCA. The feature selection methods were implemented with the top 20% of weighted attributes being retained for the model. The methods used were Information Gain, Chi-squared, and SVM. The feature selection process proved to be very encouraging with significant gains for both algorithms as can be seen in Table 5.

Model	Information Gain	Chi-squared	SVM
NB	97.67%	97.67%	97.77%
SVM	94.76%	94.76%	96.80%

Table 5: Feature selection

4.3.4 Results of preprocessing

For SVM, the optimal preprocessing techniques employed were tokenization, case transform, stopword filtering, stemming (snowball), with TF-IDF vector creation and SVM feature selection. These preprocessing techniques achieved an accuracy of 96.80%. For Naive Bayes, tokenization, case transform, stopword filtering, token filtering, n-grams, and stemming (snowball), with Binary Term Occurrence vector creation and SVM feature selection were the optimal preprocessing configuration, and achieved an accuracy of 97.77%.

4.4 Develop models

To assess the performance and development of the models, training and test data was generated with K-fold cross validation, with K=10.

4.4.1 Developing SVM

The best accuracy achieved during preprocessing for SVM was 94.47% and so the model development was carried out with this figure as a comparison. To reduce possible overfitting, it was decided to introduce a cost function. However, introducing any cost function vastly reduced the accuracy of the model, as can be seen in Table 6.

Model	Accuracy C=1	Accuracy C=0.1	Accuracy C=0.01
SVM	64.21%	65.96%	67.90%

Table 6: Cost functions

Various kernel functions were attempted and the results can be seen in Table 7. The default dot kernel function performed best.

Model	Dot	Radial	Polynomial	Neural	Anova	Epachnenikov	Gaussian	Multiquadratic
SVM	94.47%	52.57%	56.17%	50.24%	50.24%	52.57%	51.37%	50.24%

Table 7: Kernel functions

4.4.2 Developing Naive Bayes

As there are no parameters for Naive Bayes to adjust, further development of the algorithm was done through ensemble methods, as detailed in Section 4.4.3.

4.4.3 Ensemble methods

Various ensemble methods were used with either SVM or Naive Bayes as a base learner. These included AdaBoost, Stacking, Bagging, and Vote. The results are detailed in Table 8. For most ensemble methods, there were minor improvements for SVM. However, for Naive Bayes, none of the ensemble methods were able to improve on the accuracy of 97.77% for the model.

Model	AdaBoost	Stacking	Bagging	Vote
Naive Bayes	97.77%	81.57%	97.67%	96.22%
SVM	96.61%	96.03%	96.51%	96.22%

Table 8: Ensemble methods

4.5 Deployment results

Unseen patent data was gathered from the U.S patent website from the month of December, 2013. A random selection of software and non-software patents were crawled using the same methods discussed in Section 4.1. The abstracts were written to and stored in an Excel file, test_deploy.xlsx, containing 245 rows of data. The Naive Bayes model with the same preprocessing techniques, wordlist, were then deployed on the new data. The results were mixed. Of the software patents, 99.25% were predicted correctly. However, only 82.14% of non-software patents were predicted correctly.

5 Results and Discussion

The objective of this paper was to create a software patent classification model that can categorize new patents as either software or non-software.

The web mining process was straightforward and included web crawling and information extraction through Xpath expressions. The data that was gathered proved to be of good quality and did not require further data cleansing. This was confirmed through visualization of the data, such as tag clouds and phrase net, which also provided valuable insights into the nature of the data and significant words.

To preprocess the text data, several text processing techniques were tried using the baseline accuracy scores for Naive Bayes and SVMs as guidelines. If a preprocessing technique improved the accuracy of each model, it was retained; otherwise, it was discarded. Tokenization was required to create a dictionary, or bag-of-words and so the first technique to be evaluated was case transform. Although case transform lead to a very slight decline in accuracy for both algorithms, it was decided to retain this technique to remove any potential bias. In addition to the standard stopword list, a patent-specific stopword list was also used, and this lead to increased accuracy for both algorithms. There was mixed results for filtering tokens of less than two characters as it did not affect the accuracy for Naive Bayes but reduced the accuracy for SVM. It was then discarded from the SVM model. Bi-grams and tri-grams lead to a reduced accuracy for SVM, but an increase in accuracy for Naive Bayes was observed for bi-grams. Various stemming algorithms were employed and the Snowball algorithm lead to the greatest accuracy for both models. After this stage, it was determined that the optimal processing techniques for SVM were tokenization, case transform, and stopword filtering; and for Naive Bayes were tokenization, case transform, stopword filtering, token filtering, bi-gram generation, and Snowball stemming.

All word vector creation algorithms were tried for both models. For SVM, the default TF-IDF vector proved to be most accurate. However, for Naive Bayes, Binary Term Occurrence proved best. This is in line with experiments by Schneider (2004), who discovered that reducing word frequency information through the creation of a Binary Term Occurrence vector can lead to an increase in accuracy for a Naive Bayes classifier.

Both pruning and feature selection were tried to reduce the number of attributes for modeling. Pruning terms are occur in less than 3% of documents did not prove useful for either algorithm. However, feature selection methods that retained the top 20% of terms was a much more beneficial exercise. Various selection methods were tried but weighting by SVM proved the most accurate for both algorithms.

Using 10-fold cross-validation, various models were developed and tested based on the optimal preprocessing techniques that were discovered for both algorithms. The various parameters of the SVM model were modified but all changes saw reduced accuracy. As the parameters for Naive Bayes cannot be modified, further development of that model was done through ensemble methods. Ensemble methods were also tried using SVM as the base learner. Although ensemble methods did improve the accuracy of SVM, in particular the Adaboost method, it failed to deliver an increase in accuracy for Naive Bayes. In the end, it was determined that the optimal model was Naive Bayes with a Binary Term Occurrence vector and SVM feature selection.

The final model and wordlist, were written so that they could be deployed on unseen data. When the model was run, the results were mixed: 99.25% of software patents were correctly predicted, but only 82.14% of non-software patents were correctly predicted. While this could indicate that the model was over-trained, it could also be due to the fact that only a relatively small proportion of patents were analyzed. With over 8 million patents registered with the U.S. patent office (U.S. Patent Office, 2010a), the data used in this project reflects only a tiny proportion of the available data. Therefore, a representative subset of this data would have been quite large and the process of gathering that data would be too much of an undertaking for this project.

6 Conclusion

Despite patent analysis being an active area in both the academic and commercial sectors, little success has been had in text categorization of patents, at least for the categorization of patents to the correct IPC (or alternative classification body) classification. For software companies, identification of new software patents is a critical task in new product development and innovation. Patent searching is also used to identify prior art that might render a possible patent invalid. Searching for software patents, from a legal perspective, is very important to avoid possible intellectual property conflicts. In addition, it is important that patent analysts can identify which new patents are software, to avoid the considerable cost of filing duplicate or redundant patents.

Therefore, given the obvious utility of a software patent classifier, it is surprising how little work has been done in the area. Although the results from this research are mixed, it is felt that a more comprehensive and representative training data set would lead to a more reliable software patent identification system. Such a system would be straightforward to implement in a business setting and would easily bring tangible commercial benefits.

References

Bessen, J. and Hunt, R. (2004). An empirical look at software patents. *Research on Innovation*.

Fattori, M., Pedrazzi, G., and Turra, R. (2003). Text mining applied to patent mapping: a practical business case. *World Patent Information*, 25:335–342.

Graham, S. and Mowery, D. (2003). *Patents in the Knowledge-Based Economy*. The National Academies Press.

Guadamuz, A. (2006). The software patent debate. *Journal of Intellectual Property Law and Practice*, 1(3):196–206.

Hall, B. and MacGarvie, M. (2006). The private value of software patents. NBER Working Paper 12195.

IBM (2013). Many eyes. viewed 17 December 2013.

Kaur, M. and Sapra, R. (2013). Classification of patents by using the text mining approach based on pca and logistics. *International Journal of Engineering and Advanced Technology*, 2(4):711–714.

Krier, M. and Zacca, F. (2002). Automatic categorisation applications at the european patent office. *World Patent Information*, 24(3):187 – 196.

Larkey, L. (1998). Some issues in the automatic classification of U.S. patents. In *Learning from text classification. Papers from the 1998 workshop*, pages 87–90. AAI Press.

Larkey, L. (1999). A patent search and classification system. In *Proceedings of the fourth ACM conference on digital libraries*, DL 99, pages 179–187, New York, NY, USA. ACM.

Layne-Farrar, A. (2005). Defining software patents: A research field guide. Working Paper.

Liang, Y. and Tan, R. (2007). A text-mining-based patent analysis in product innovative process. In Leon-Rovira, N., Editor, *Trends in computer aided innovation*, Volume 250 of *IFIP The International Federation for Information Processing*, pages 89–96. Springer US.

Macher, J., Mowery, D., Industry, C., Board on Science, T., Affairs, P., and Council, N. (2008). *Innovation in global industries: U.S. firms competing in a new world (Collected Studies)*. National Academies Press.

Miner, G., Elder IV, J., Hill, T., Nisbet, R., Delen, D., and Fast, A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*, chapter Text classification and categorization. Academic Press.

Poltorak, A. and Lerner, P. (2002). *Essentials of intellectual property*. John Wiley and Sons.

Potter, K. and Hatton, R. (2013). Data mining of u.s. patents: Research trends of major technology companies. In *SAS global forum 2013*.

PWC (2013). PWC global 100 software leaders. Technical report, PWC.

Schneider, K. (2004). On word frequency information and negative evidence in naive bayes text classification. In *Natural language processing, ESTAL*.

Tseng, Y., Lin, C., and Lin, Y. (2007). Text mining techniques for patent analysis. *Information Processing and Management*, 43(5):1216 – 1247.

Tseng, Y., Wang, Y., Juang, D., and Lin, C. (2005). Text mining for patent map analysis. In *IACIS pacific 2005 conference proceedings*, pages 1109–1116.

U.S. Patent Office (2010a). U.S. patent database. viewed 15 December 2013.

U.S. Patent Office (2010b). U.S. patent stopwords. viewed 18 December 2013.

World Intellectual Property Organization (2013). International patent classification (IPC) official publication. viewed 15 December 2013.

Wu, C., Ken, Y., and Huang, T. (2010). Patent classification system using a new hybrid genetic algorithm support vector machine. *Applied Soft Computing*, 10(4):1164 – 1177.

Zhang, X. (2014). Interactive patent classification based on multi-classifier fusion and active learning. *Neurocomputing*, 127(0):200 – 205.

Sentiment Versus Polarity within Tokens and Sentences

John Ryan [1], Markus Hofmann [2]
Institute of Technology Blanchardstown
Dublin, Ireland
[1] leabharnua@gmail.com
[2] markus.hofmann@itb.ie

Abstract

This research paper demonstrates the possibility of registering the dominant polarity of each sentence within a document, as opposed to understanding the sentiment of individual tokens, by using existing sentiment libraries. The objective is to demonstrate clarity of results that individual tokens may lack. As a control measure, this research paper compares two results from the same extracted text; the first avails of standard sentiment analysis on tokens and the second regarding processes that are applied to retrieve polarity from individual phrases. This paper demonstrates that when analysing sentences, polarity is more effective than defining a specific emotion; while sentiment is beneficial, more clarity can be obtained through polarity. It also indicates that polarity scoring of sentences provides an equally worthwhile method of understanding the corpus when compared to using alternative systems that rate the whole document.

Keywords: Text Analysis, Sentiment, Polarity

1 Introduction

Applying sentiment or polarity analysis to individual tokens and then adding those results together to provide an aggregate positive and negative rating for a whole piece of text lacks the nuanced visibility of understanding the overarching polarity from each of the document's sentences. The following sections explore available technologies and methods to understand if there are standard accessible tools to achieve this sort of multi-layered sentiment analysis. Section 2.1 analyses at a token level while Section 2.2 explores at a phrase level. As a corpus, this research paper avails of the political reactions to the `Commission of Investigation Report in the Catholic Diocese of Cloyne`, which was published in 2011. These data transcripts were obtained from `http://debates.oireachtas.ie/XML/D/2011/DAL20110720.XML`.

Bias needs to be carefully considered as 35.86% of words are rated `joy` within the dictionary used for this analysis. To also reduce bias, term occurrences ensure only one instance of a word is analysed per phrase within Section 2.2. Since the document for this research is negative, there is also the issue of human bias; a subjective opinion can therefore be contrasted with the objective results. Throughout the following sections, sentence fragments are quoted to demonstrate how the objectively emotion labelled term resides within the context of the phrase to show whether it is subjectively appropriate.

The analytic software used to extract information from the political corpus was the open source data mining suite RapidMiner (`http://rapidminer.com`). Polarity Analysis was applied using R software with a sentiment package (Jurka, 2012c). Phrase level results were produced through Microsoft Excel. The Wordclouds in this section were created within the online application `http://www.wordle.net/`, but could have also been generated using IBM's service ManyEyes. For expediency purposes in this paper; the abbreviation for an Irish politician is TD (Teachta Dála) and the `Commission of Investigation Report in the Catholic Diocese of Cloyne` is `Cloyne Report`.

2 Applying Sentiment and Polarity Methods

2.1 Standard Sentiment Analysis on Tokens

This section focuses on identifying both emotions and polarity for single words (tokens) in a document using a sentiment package within R software; sourced from Jurka (2012c). Sentiment analysis facilitates the identification of opinions expressed through NLP (Natural Language Processing). There are 3 sentiment analysis measurements:

1. Polarity: According to Shelke et al. (2012) these can be classed as positive (i.e. satisfied, happy), negative (dissatisfied, disappointed), neutral i.e. no opinion (generally when just stating facts).

2. According to Strapparava and Mihalcea (2008) there are 6 main emotions that can be identified within text; `Anger`, `Disgust`, `Joy`, `Surprise`, `Sadness` and `Fear`. This is corroborated by the sentiment package used in R software (Jurka, 2012c). Ways in which to calculate emotions can be through the Wordnet-Affect system where words are annotated by emotions sourced within the Wordnet-Affect Lexicon (Strapparava and Mihalcea, 2008).

3. Best-fit: this is a measurement within the sentiment package in R, that indicates the most likely emotional category in which the text opinion could be classified (Jurka, 2012c).

Analysing the Taoiseach Enda Kenny's speech in response to the release of the `Cloyne Report` on 20th July 2011, indicates `sadness` (Table 1) as one of the primary emotions. The dominant rating of `joy` (51.55) should be treated with caution and requires human subjective analysis to determine if it is valid. After subjectively reviewing, and considering the level of bias within `joy`, that emotion's score can be challenged; it shows clear conflicts between the classified words and the context (Table 2); for example `pick` relates to "it could take the victims and their families a lifetime to pick up the pieces", and `comfort` relates to "little that I or anyone else in the House can say to comfort that victim". However, there does not appear to be classification conflicts with negative emotions:

1. "WORD: abhorrence CAT: anger SCORE: 5.87211778947542"

2. "WORD: dismay CAT: fear SCORE: 5.27299955856375"

3. "WORD: dismay CAT: sadness SCORE: 5.61312810638807"

4. "WORD: frustrate CAT: anger SCORE: 5.87211778947542"

5. "WORD: heartbreaking CAT: sadness SCORE: 5.61312810638807"

6. "WORD: suffering CAT: sadness SCORE: 5.61312810638807"

Table 1: Sentiment Analysis of the Taoiseach's response to the Cloyne Report

ANGER	DISGUST	FEAR	JOY	SADNESS	SURPRISE	BEST_FIT
13.21	3.09	7.34	51.55	18.57	2.79	"Joy"

Thus, `sadness` is registered as the predominant emotion with a value of 18.57, followed by `anger` with a value of 13.21. Words such as `dismay` are categorised as both `fear` and `sadness`, `abhorrence` as `anger` and `suffering` and `heartbreaking` as `sadness`. The objective scores accurately reflect the subjective sentiment of the piece. The overall polarity best-fit rating is `neutral`, Table 3, with the positive (365.078) and negative (328.82) ratings also registering as quite close in value. This `neutral` recommendation suggests that the statement was factual in nature, while it was that, it was

Table 2: Enda Kenny TD: Joy Classification for the Cloyne Report

1	"WORD: live CAT: joy SCORE: 6.31535800152233"
1	"WORD: pick CAT: joy SCORE: 6.31535800152233"
1	"WORD: protect CAT: joy SCORE: 6.31535800152233"
1	"WORD: regard CAT: joy SCORE: 6.31535800152233"
1	"WORD: approval CAT: joy SCORE: 6.31535800152233"
1	"WORD: comfort CAT: joy SCORE: 6.31535800152233"
1	"WORD: heart CAT: joy SCORE: 6.31535800152233"
1	"WORD: sympathy CAT: joy SCORE: 6.31535800152233"

also laced with negative emotions as discovered through sentiment analysis. Thus, it is worthwhile further investigating the way various words were marked, as can be seen from this short extract from the overall output report (Table 4). Table 4 also indicates why the positive polarity rating was so high and shows how meanings can be misconstrued within the produced results.

Table 3: Polarity Analysis of the Taoiseach's response to the Cloyne Report

POS	NEG	POS/NEG	BEST_FIT
365.078	328.32	1.11	"neutral"

For example the words `sanity` and `humanity` are rated as positive but reading the context, in which they originated from, seems to portray the opposite. The Taoiseach is stating that priests are "struggling to keep their humanity, even their sanity". The word `truth` while positive in most regards, again taken within the context where the Taoiseach is saying that "nothing could be further from the truth" with regards to what the Catholic Church stood for; "humility and compassion". When compared to the overall response from the Dáil, this would be TDs from all types of political backgrounds including opposition party leaders such as Micheál Martin (FF) and Independent TDs such as Mick Wallace, the overriding categorisation, outside of `joy`, is `sadness` followed by `fear` (Table 5). The overall polarity is rated as `negative` (Table 6). The polarity dictionary contains 64.05 % negative terms, so there is a caution of bias. However, there is a unity in both the attitude and response to the findings of the `Cloyne Report`. Again `joy` is discounted because of the use of words such as `respect`, `move`, `pleased`, `proud`, `regard`, `wall`. What is far more revealing are the negative words used, as referenced in Figure 1. As for polarity, strongly subjective negative words including `worthlessness`, `tragically` and `wretched` help to influence the scoring.

2.2 Sentiment Analysis using Phrases

Section 2.1 indicates that `joy` was discounted after judging its context to be inappropriate. This leads to an interesting conclusion; tokens taken on their own can result in questioning how those words reside in the overall context of a sentence, thus taken as a whole instead of being an independent entity. For example, "The Great Depression", according to the Wordnet-Affect Lexicon dictionary that is used within the sentiment package in R, `the` is eliminated as a stopword, `great` is `joy`, `depression` is `sadness`, so what does that mean for "The Great Depression"? Polarity rather than sentiment is proven in this section to be more effective; for example, Table 7 shows that negative emotions can be applied to

Table 4: Enda Kenny TD: Classification for the Cloyne Report Output Extract

1 ”WORD: heart CAT: positive POL: weaksubj SCORE: 7.7510451179718”
1 ”WORD: heartbreaking CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: holy CAT: positive POL: weaksubj SCORE: 7.7510451179718”
1 ”WORD: horrors CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: humanity CAT: positive POL: weaksubj SCORE: 7.7510451179718”
1 ”WORD: humiliation CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: humility CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: sanity CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: sensitivity CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: suffering CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: suspicions CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: sympathy CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: thankfully CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: torture CAT: negative POL: weaksubj SCORE: 8.33686963728496”
1 ”WORD: truth CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: unable CAT: negative POL: weaksubj SCORE: 8.33686963728496”
1 ”WORD: unprecedented CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: unwilling CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: uphold CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: victim CAT: negative POL: strongsubj SCORE: 9.0300168178449”
1 ”WORD: vulnerable CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: wish CAT: positive POL: strongsubj SCORE: 8.44419229853175”
1 ”WORD: world CAT: positive POL: strongsubj SCORE: 8.44419229853175”

Figure 1: TD Negative Response (fear, sadness, disgust, anger) to the Cloyne Report

Table 5: Sentiment Analysis of TDs' response to the Cloyne Report

ANGER	DISGUST	FEAR	JOY	SADNESS	SURPRISE	BEST_FIT
95.42	37.08	112.80	272.59	147.67	21.00	"Joy"

Table 6: Polarity Analysis of TDs' response to the Cloyne Report

POS	NEG	POS/NEG	BEST_FIT
2557.93	3098.63	0.83	"negative"

the same word, however with polarity they would all be registered as negative. Also, as evident in this section, there is a more substantial dictionary available for polarity. Research by Wilson et al. (2009), who analysed documents at a phrase level, clearly demonstrated that the context of polarity is an important consideration. This section tries to achieve this by highlighting the construction of a polarity sentence structure analysis model. Developing this process would allow each word to be tokenized not as if they belonged to an overall document, but as if they actually belong to an individual sentence; thus acting as a mini-document within the whole document. The number of negative and positive words within a sentence could be extracted and analysed using this type of model. For example, if there are 3 negative emotion type words in a sentence, and 2 positive type words in the same sentence, then that sentence could be judged to be in the majority a negative emotion type sentence and the strongest identified emotion is the number of classifications it receives within the sentence.

A number of procedures had to be undertaken in order to create this model:

1. The same WordNet-Affect Lexicon that is used within the Sentiment package in R had to be downloaded (Jurka, 2012a). This csv formatted file is imported into the main spreadsheet application that is used to process the majority of the sentiment analysis calculations.

2. Two processes within RapidMiner had to be constructed. The first to break the corpus into sentences (Figure 2), and the second to tokenize and write the results to a spreadsheet (Figure 3). The 1st process (Figure 2) takes the corpus, removes the html tags within it, breaks the text into sentences and saves them into a file called "sentencemaker.xls". 31 sentences were produced for the Taoiseach's response to the `Cloyne Report` (1st sentence split into 2 phrases). The 2nd process takes these sentences, extracts their tokens and counts their term occurrences. The most crucial step in this process is the assigning of a unique tag to the tokens using the `GenerateID` operator, so that they can be grouped by their own sentence. They are then written to "resultset.xls".

3. At this stage, all tokens are assigned an id representing their sentence origin and are stored within Excel. The WordNet-Affect Lexicon (from Step 1) is then imported into a new worksheet (Name:`EmotionStem`) within this file; an attribute is inserted for column A that contains the first 5 characters of the main word in column B (Table 7), thus creating a 5 character stem of that word. This is in order to maximise a beneficial outcome in matching, as the stemmed word is compared to the first 5 letters of the token.

4. The next task is that the results are transposed in a new worksheet (Name:`Transposed`) so that

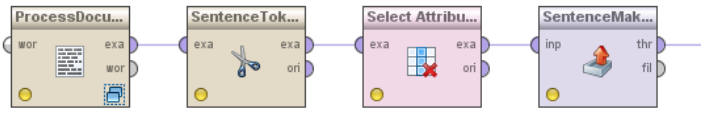


Figure 2: Sentiment Analysis Model Step 1

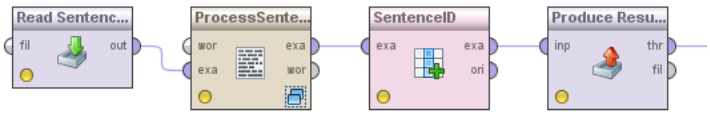


Figure 3: Sentiment Analysis Model Step 2

B144 =IF(Transposed!B144>=1, (MATCH(LEFT($A$144,5),EmotionsStem!$A$1:$A$1542,0)),0)

	A	B	C	D	E	F	G	H	I	J	K
1	Words	Sent-1	Sent-2	Sent-3	Sent-4	Sent-5	Sent-6	Sent-7	Sent-8	Sent-9	Sent-10
140	complete	0	0	0	0	0	0	0	0	0	0
141	compliance	278	0	0	0	0	0	0	0	0	0
142	comprise	0	0	0	0	0	0	0	0	0	0
143	concerning	284	0	0	0	0	0	0	0	0	0
144	condition	0	0	0	0	0	0	0	0	0	0

Figure 4: Sentiment Spreadsheet- Matching Tokens to WordNet-Affect Lexicon List

the words are in column A and the sentences (relabelled as Sentence-1, Sentence-2...) become the column attributes. A new worksheet (Name:`Match`) needs to be created within the application (words again as the column A row headers and the sentences as the column attribute headers), so that the tokens can be matched to the lexicon dictionary in order to understand if they can be categorised under an emotion-type. The formula used for each cell is detailed below:

```
=IF(Transposed!B2>=1, (MATCH(LEFT($A$2,5),
EmotionsStem!$A$1:$A$1542,0)),0)
```

Thus if the token has a term occurrence value present within the cell, then it undergoes matching to a Stemmer version of the lexicon dictionary (within `EmotionStem`). If a resulting match is found (first position in the range), then the actual row number is recorded. So if the word `Affection` (stem: affec) was in row 23, and it matched a token, then the number 23 is recorded (Figure 4).

Table 7: Snapshot of Emotions List obtained from WordNet-Affect Lexicon

Stem (5)	Original	Emotion
abhor	abhorr	anger
abhor	abhorr	disgust
abhor	abhorrent	disgust

5. Once the `Match` worksheet is completely populated, the next stage is to create a new worksheet (Name:`Emotions`) within the application that includes the original lexicon dictionary which excludes the stem (Figure 5). Underneath that emotions categorisation list, a new table is created (populated from the `Match` worksheet). The same structure applies as the rest of the result-set tables, namely words in the A column and sentences as attribute column headers. The cells are populated by the formula detailed below:

```
=IF(ISNUMBER(Match!B2)=TRUE,Match!B2,0)
```

Essentially this cleans the results presentation; if there is a value then it is displayed, otherwise the value zero is recorded. Underneath this table, on the same sheet, a final table is created, one with the same words in the A column and sentences as attribute headers. In this area, the row position numbers (which were gathered in Step 3) become the actual emotions that they describe. In order to accomplish this, each cell is populated via a formula that is detailed below:

```
=IF(B1549>0,HLOOKUP(Emotions!$B:$B,$B$1:$B$1542,B1549),0)
```

	A	B	C	D
1	abhor	anger	1	17
2	abhorr	anger	2	
3	abhorr	disgust	3	
4	abhorrence	anger	4	
5	abhorrent	disgust	5	
6	abomin	anger	6	
7	abomin	disgust	7	
8	abominably	disgust	8	
9	abominate	anger	9	
10	abomination	anger	10	

Figure 5: Layout of Emotions Categorisation Within Spreadsheet

=IF(B1709>0,HLOOKUP(Emotions!$B:$B,$B$1:$B$1542,B1709),0)

H	I	J	K	L	M
Words	Sent-1	Sent-2	Sent-3	Sent-4	Sent-5
democratic	0	0	0	0	0
deplores	sadness	0	0	0	0
describes	0	0	0	0	0
deserve	0	0	0	0	0

Figure 6: Sentiment Spreadsheet- Labelling Emotions Depending on Cell Number

For this to work, the word dictionary needs to display all words within the A column (row numbers A1 to A1542) and the emotions must be listed in the B column (row numbers B1 to B1542). An example of the layout is represented in Figure 5. The formula takes the value of the number, if there is one, and finds the emotion label that is associated with that row (Figure 6).

6. After the emotion categorisation table has been completed, the actual results can be calculated, and this is where the breakdown of the sentences can be undertaken. Thus, each of the 31 sentences undergoes a series of 6 checks (Table 8). In summary, the sentence construction can now be fully explored.

Breaking the first phrase, ”That Dáil Éireann...the Criminal Justice”, into emotions (Figure 7) indicates a better understanding of how the classifications are achieved using the previously outlined steps within this model.

Table 8: Example of Results Output per Sentence for Sentiment Analysis Model

	Sentence1
Fear	=COUNTIF(I2:I257, ”fear”)
Sadness	=COUNTIF(I2:I257, ”sadness”)
Disgust	=COUNTIF(I2:I257, ”disgust”)
Surprise	=COUNTIF(I2:I257, ”surprise”)
Joy	=COUNTIF(I2:I257, ”joy”)
Anger	=COUNTIF(I2:I257, ”anger”)

Objective results indicate that the primary emotion in the 5th sentence is negative, reflecting `anger`; ”However, the Cloyne report has proved to be of a different order because for the first time in this country a report on child sexual abuse exposes an attempt by the Holy See to frustrate an inquiry in a sovereign, democratic republic as little as three years ago, not three decades ago.”. While the 6th sentence elicits `sadness`, ”In doing so the report excavates the dysfunction, disconnection and elitism that dominate the culture of the Vatican to this day.” Sentence 22, ”The Tánaiste left the archbishop clear on two things:

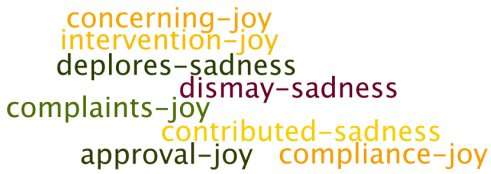


Figure 7: Breakdown of Emotions within Sentence 1

the gravity of the actions and attitude of the Holy See and Irelands complete rejection and abhorrence of same." is rated as both `anger` with the word `abhorrence` and `joy` for the word `complete`. Again, by reading the sentences and understanding the tokens as they are within the sentence, one can subjectively judge whether `complete` is anything but `joy` in this regard. Thus, this model allows the flexibility to understand the classifications and where to question them.

The polarity measurement can also be applied to this same model by substituting the emotions classifications with the polarity classifications; downloaded from Jurka (2012b). There are over 5 times the amount of words within this dictionary than the emotions lexicon (Table 9). This factor is highlighted in the subsequent results with a more detailed matching and analysis.

Table 9: Snapshot of Polarity List

Stem (5)	**Original**	**Polarity**
aband	abandoned	negative
aband	abandonment	negative
aband	abandon	negative
abhor	abhorred	negative

Overall, using the same count-if formula as was used for counting emotions (Table 10), a new count-if counts `positive` and `negative` occurrences which informs that sentence 1 from the `Cloyne Report` is indeed negative; referenced from Figure 8 with a negative rating of 23 instances versus a positive of 11. This type of analysis is calculated throughout the 31 sentences; summarised

```
=IF(I260=I259, "NEUT", IF(I260>I259,"NEG","POS"))
```

with 10 sentences registered as negative, 11 as positive and 10 that are split 50/50 so could be either Positive or Negative and require further analysis; an advantage of text mining is that the corpus in question can be read and opinions applied subjectively.

Table 10: Example of Results Output per Sentence for Polarity Analysis Model

	Sentence1
Positive	=COUNTIF(I2:I257, "positive")
Negative	=COUNTIF(I2:I257, "negative")

There is one caveat when using this model. A 5 character Stemmer could result in erroneous results, such as `complaints` registered as `joy` (Figure 7), due to the fact that `compl` is actually `complacence` (`joy`) within the R sentiment package. Overall though, the majority of the words supersede that and provide a correct outcome (which is one of the reasons to judge the whole sentence as opposed to individual word terms). As previously stated within this research paper; the rating of `joy` has to be monitored closely in any result to ensure it does not skew the actual sentiment interpretation. Other Stemmer lengths could also be tested.

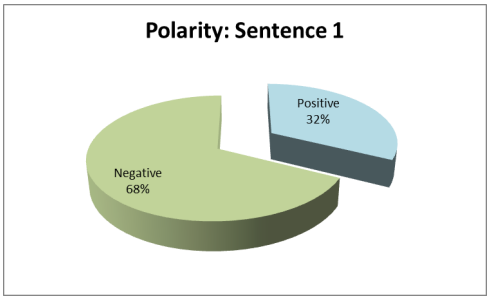


Figure 8: Polarity within Sentence 1

3 Conclusion

Documents can be rated for polarity by their phrases. This non-trivial manual process involves downloading and formatting specific sentiment libraries or dictionaries, breaking the extracted text into sentences and assigning a unique identifier per sentence. This paper has demonstrated that due to the larger range of recorded results for polarity, a more comprehensive result definition can be generated. Polarity also has a much better clarity of results; where the word is either positive or negative and not a range of feelings where two emotions may be based on the same token. For example comparing Table 7 with Table 9 shows the difference and the benefits of using polarity over sentiment, where all versions of `abhorr` are simply registered as negative within the polarity dictionary. Another potential avenue to explore would be to emphasise the level of negativity or positivity of a word depending on the number of times it appears within a sentence. The document length was short, thus further work could include larger files with a more varied subjective outcome (it was recognised that the piece used was negative in sentiment). What this research paper proves explicitly, is that there is merit to exploring not just tokens as an individual entity but as part of a sentence structure (both systems can be used in a complimentary fashion); to fully understand if a phrase is indeed negative or positive. Further work could be refining the model developed, adjusting the Stemmer lengths as referenced in Section 2.2 and applying this analysis to other text extracts to understand if the formulas devised for the extracted document within this research paper could be replicated for other corpus. It is only the beginning of a conversation, not a finite conclusion.

References

Jurka, T. P. (2012a). Sentiment r package emotions: `https://github.com/timjurka/sentiment/blob/master/sentiment/data/emotions.csv.gz`. last accessed: 23 feb 2013.

Jurka, T. P. (2012b). Sentiment r package subjectivity: `https://github.com/timjurka/sentiment/blob/master/sentiment/data/subjectivity.csv.gz`. last accessed: 23 feb 2013.

Jurka, T. P. (2012c). Sentiment: Tools for sentiment analysis: `http://cran.open-source-solution.org/web/packages/sentiment/`. last accessed: 17 feb 2013.

Shelke, N. M., Deshpande, S., and Thakre, V. (2012). Survey of techniques for opinion mining. *International Journal*, 57(13):30–35.

Strapparava, C. and Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM.

Wilson, T., Wiebe, J., and Hoffmann, P. (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433.

Automated Detection of Cyberbullying

David Colton [1], Markus Hofmann [2]

[1] Institute of Technology Blanchardstown (ITB), Dublin 15
davidcolton@gmail.com

[2] Institute of Technology Blanchardstown (ITB), Dublin 15
Markus.Hofmann@itb.ie

Abstract

With the advent of instant messaging, chat rooms and social media websites, the proliferation of the internet onto smart phones and tablets and the anonymity these services and devices offer, the bully has moved on-line and now has 24 hour access to their victims. Unfortunately the sometimes tragic consequences of cyberbullying are plain to see with the victims suffering from stress, lack of sleep and in extreme cases mental health issues and suicidal thoughts. In a perfect world the victim of this bullying would be protected from abuse of this type. This paper undertakes an examination of Ask.fm, a site recently in the news because of its suggested link to several tragic suicides, to determine if there is cyberbullying on this site and if so whether a predictive model can be developed to support detection of bullying in a more automated manner.

Keywords: Cyberbullying, Data Mining, Text Mining, Web Content Mining

1 Introduction

Bullying is not something new. Whether in the school yard, in the street or park when playing with friends, when out socialising or in the office, bullying is something nearly everyone will have had either a direct or indirect experience of during their lives. Although never a pleasant experience the target of the bullying had the power to remove themselves from the source of their anguish. Now, however, with the advent of instant messaging and social media, the bully has moved on-line and has 24 hour access to their victims. In this modern age of digital communications it is becoming increasingly difficult to function without a digital footprint making it impossible for the victim to every fully escape and feel safe from the reaches of a bully. Unfortunately the sometimes tragic consequences of cyberbullying are plain to see with headlines like "Cyber-bullies claimed lives of five teens" [8] depressingly becoming an all too frequent occurrence.

Ask.fm, a social networking website that uses a question and answer format to allow its users interact, is a social media site which has recently featured predominately in the news headlines with multiple teenage suicides attributed to the trolling and cyberbullying the victims have suffered on the site. Based in Riga, Latvia, the sites ease of access and the anonymity offered means that it is increasingly being used as a means to communicate abusive, bullying and sexualised content [11].

In this paper an approach is proposed that will screen questions on the Ask.fm site to determine if they contain text that could be deemed as cyberbullying. In the next section, Background, an introduction to the Ask.fm site is given. Following this is a Literature Review to determine if there are existing published papers that describe approaches that have shown success in detecting cyberbullying. A brief introduction to the data is provided before the Methodology section describes the modelling performed on the Ask.fm data. The Results and Discussion section provides details on how each of the various models and learners performed before the paper is completed with conclusions and future work.

2 Background

Ask.fm is a social networking website that uses a question and answer format to allow its users to interact. In order to better understand how to detect cyberbullying posts on Ask.fm it is important to understand how the site is constructed and used by its members.

Ask.fm provides a standard new account experience to users when compared to other social media sites. When creating an account the user is asked to provide typical account information such as the user id to be used, the users name, email etc. When a new account has been created probably the most important thing for the new user to do is to configure their privacy settings. If a user decides to Allow anonymous questions then anyone, even people who do not have accounts on Ask.fm, can ask anonymous questions. Questions can be answered using text, pictures, videos or a combination. If a user decides not to answer a question it can easily be deleted or if a user is continually receiving abusive and bullying posts the asker of the question(s) can be blocked.

3 Literature Review

In the literature there any many discussions about what constitutes cyberbullying and the roles and forms that cyberbullying takes. Three main labels into which cyberbullying can be categorised are described by Dinakar et al. [4] as sexual, sexist attacks typically against women or sexual minorities (for example homophobic), racial or cultural attacks, racial slurs or attacks against a cultural minority and its traditions and intelligence and physical appearance attacks which are direct attacks against a persons mental capacity or, for example, their weight, height or appearance. The U.S. Department of Health and Human Services (DOH) [9] reinforce these ideas and describe bullying as the unwanted, aggressive and repeated behaviour, typically amongst school aged children, where there is a perceived or actual imbalance of power. Three types of behaviours described are verbal bullying, for example name calling or teasing, social bullying like exclusion or spreading rumours and physical bullying which is actual or threatened violence. The DOH say that cyberbullying is bullying using communication tools such as instant messaging, chat sites and socials networks using smart phones, computers or tablets and highlights that cyberbullying can happen 24 hours a day 7 days a week. Applying these definitions of what constitutes cyberbullying to the sample questions dataset it was easy to classify the questions as either bullying or not bullying.

A common approach for the first initial steps in detecting cyberbullying text in social media sites was lexical analysis using a weighted Term Frequency Inverse Document Frequency (TF-IDF) scheme. At the simplest feature level a message is represented as a vector of TF-IDF values for each term in the post. Chen et al. [1] suggests that this bag of words approach will give low accuracy in subtle offensive language detection and suffers from a high false positive in heated arguments or exchanges between friends and that an N-gram approach can give better results with Bi-grams and Tri-grams commonly used. Nahar et al. [10] supplements a TF-IDF scheme with Latent Dirichlet Allocation (LDA), a topic-modelling approach, which is used to understand the underlying semantic topics (for example bullying). Yin et al. [12] suggested that context could be used to help identify bullying texts along with a TF-IDF vector of terms. Following a review of their data it was discovered that foul language, especially when used in conjunction with second person pronouns such as "you" or "yourself" was indicative of harassment. A library of sentence formats such as "Pronoun $\cdots$ Foul Word $\cdots$" or "I Foul Word Pronoun $\cdots$" was constructed to perform sentiment analysis on the text.

Finally, Chen et al. [1] identifies that the text used in posts on social media sites is usually highly unstructured and informal where normal grammar rules and spelling are not used. The use of slang words and emoticons also exasperates the situation. Dinakar et al. [5], Kontostathis et al. [6], Kontostathis et al. [7] and Dadvar et al. [3] all lament the lack of good quality annotated datasets with many groups having to create their own by scraping content from various sites. These are issues that also affected this paper.

4 Data

A lot of data was scraped from Ask.fm including 85,000 questions and answers and from this rather large dataset a simpler, smaller more manageable dataset of approximately 8,500 sample questions were extracted for use in this paper. Each of the questions in this dataset was then classified as either bullying or not bullying. In total 1,175 questions were classified as bullying.

An initial brief exploration of the data showed that two thirds of the 8,483 sample questions were asked anonymously and of the questions that were classified as bullying 85%, were asked anonymously. Comparing the lengths of bullying questions to non-bullying questions showed that the former were slightly shorter at 44 characters whilst the latter were 48 characters. Of the 15% of bullying posts that were not asked anonymously, 175, there were very few repeat offenders with only 8 users asking more than 1 bullying questions and of these 8 users 7 of them asked all their abusive questions to the same user. Words clouds of both classes were created and interestingly some words, like, love and just, were prominent in both.

5 Methodology

In this section the creation of a model to predict whether a piece of text can be classified as bullying or not is described. All mining was performed in RapidMiner. To begin, a basic text mining bag of words model was developed to give an indication of the types of accuracy that may be achievable. This model started by reading the classified sample questions dataset from a database and then splitting the data into a training set, 80%, and a testing set, 20%, using stratified sampling. After converting the nominal data read from the database to text, the bag of words TF-IDF vector was created using a Process Documents from Data operator.

The contents of the Process Documents from Data operator used typical steps and included a transform case operator to convert all text to lower case, a tokenise operator to split the text of the questions into a sequence of tokens, initially set to split on non-letters, a filter stop words operator to remove English stop words from the word vector, a Porter stem operator and finally a filter tokens by length operator which was initially set to filter out words less than 3 characters and more than 25 characters.

Cross validation was performed in order to estimate the statistical performance of a learning operator and was initially set to use five validations. A simple Naive Bayes learner was chosen and then the model generated was applied and the performance of the model was estimated. With an accuracy of just under 50%, a precision of 17% and a recall of 90% these figures are not great.

5.1 Accuracy, Precision and Recall

The performance of the first simple model described above was given using three measures; accuracy, precision and recall. When evaluating the performance of the model in predicting whether a question was bullying or not precision and recall will be used. When predicting whether a question is bullying and abusive there are four possible outcomes:

1. **True Positive**: A question is predicted as bullying and is also classified as bullying
2. **False Positive**: A question is predicted as bullying but is classified as not bullying
3. **False Negative**: A question is predicted as not bullying but is classified as bullying
4. **True Negative**: A question is predicted as not bullying and is also classified as not bullying

Accuracy, Precision and Recall are calculated as:

$$Accuracy = \frac{NumberofTruePositives + NumberofTrueNegatives}{TotalNumberofExamples} \tag{1}$$

$$Precision = \frac{NumberofTruePositives}{NumberofTruePositives + NumberofFalsePositives} \tag{2}$$

$$Recall = \frac{NumberofTruePositives}{NumberofTruePositives + NumberofFalseNegatives} \tag{3}$$

5.2 Class Imbalance

The first most obvious issue with the simple model proposed earlier is that there is no allowance for the class imbalance between not bullying questions and bullying questions. A sample operator was added to the main process to select only 940 samples of each class type. The process was run again and an immediate improvement in the performance of the model was seen with the overall performance of the model increasing to nearly 64% and whilst the recall value has fallen slightly to 82% the increase in the precision of the model to 60% more than compensates. In a way this result is slightly surprising as it is suggested that Naive Bayes can handle class imbalances.

5.3 Process Documents

When reviewing the word vector it was clear that there was scope for improvement. The word vector created from the original process had stop words removed and word stemming performed. It was apparent that the resulting set of words had lost a lot of their impact and their intent. It was decided to remove both the stop words and stemming operators and replace them with an N-Gram (Terms) operator. Using the n-gram operator with a setting of 3 has the effect of merging three tokens into a single token. A lot of expressions with predictive power were created using this technique. Due to the bad language the authors have decided not to list example 3-grams.

These changes have once again had a positive impact on both the accuracy and the precision of the model with accuracy rising to 72% and precision to 74%. This rise in precision an accuracy has come at a cost though and recall has once again fallen, this time to just under 68%.

5.4 Different Learners

Having achieved relatively good results after making some minor changes to the model the next step was to change the learners used to see if further improvements can made.

The first learner chosen to replace the Naive Bayes operator is the K-NN operator. Using a Cosine Similarity measure and a value of k of 8, 9 or 10 an overall improved accuracy of around 74% was achieved however this was at a cost. Using the Naive Bayes learner the execution time for the model using 1,880 samples was in the region of 35 to 40 seconds but the execution time for the K-NN model was greatly increased and was in the region of 210 to 240 seconds which is representative of a 600% increase in processing time for a performance gain of less than 3%. Of the three values of k 9 gave the best rounded result with accuracy, precision and recall values of 73.88%, 74.22% and 73.19% respectively.

The next learner to be examined was the LibSVM Support Vector Machine (SVM) operator. Using the C-SVC SVM with a sigmoid kernel type an across the board improvement over the K-NN learner was achieved but more importantly this learner proved to be even faster than the Naive Bayes learner taking under 20 seconds, showing a 50% processing time improvement.

The next approach to improve the performance of the model was to apply feature weighting and then to just select the top n% features and pass these to the learners previously explored. The Naive Bayes, K-NN and SVM models were revisited to determine if further improvements could be made. The weight by SVM operator calculates the relevance of the attributes by computing for each attribute of the input example set the weight with respect to the class attribute. The select by weight operator selects only those attributes of an input example set whose weights satisfy the specified criterion with respect to the input weights. Given that there are approximately 12,750 attributes in the input example set, selecting only the top 10% was found to give increased model performance.

Examining the results of these three new models proved very interesting indeed. The Naive Bayes model improved the most showing nearly 20% improvement in overall accuracy from 72% to 86% with precision increasing by just under 23% to 92% and recall improving to 80% showing an 18% improvement. An additional side benefit of filtering out 90% of the example attributes was a faster model execution for Naive Bayes of just 17 seconds, a 50% improvement. Whilst the K-NN model also improved the gain was less spectacular with a modest 4.5% increase in overall model accuracy and surprisingly the overall accuracy of the SVM model actually decrease by over 15% from 77.71% to 65.64%. On the plus side thought the SVM model gave the highest bullying precision value so far at 96.23%. With this increase in precision came a spectacular decrease in the recall accuracy meaning that less questions were predicted as bullying but those that were predicted as bullying were most likely correctly predicted as so. A summary of the performance data from all the models is discussed later in this paper.

The final operator to be examined was the Singular Value Decomposition (SVD) operator. With over 12,750 attributes and nearly 1900 example the resulting matrix produced has nearly 23,000,000 values and most of these would have a zero value. This is what could be considered a sparse matrix so it was assumed that SVD would perform well on this data. Whilst the SVD model gave a very good recall result its precision and overall accuracy was low compared to the previously examined learners.

5.5 Testing the Model

Now that the optimum models have been developed they need to be tested using the data withheld during training. To do this the model was updated and the final model is shown in Figure 1. In Figure 1 the top main frame shows the complete model, in all cases the numbers represent the order in which the operators are executed. The second frame shows the full contents of operator 1 in the main frame. This sub process prepares the data by reading from the database and splitting into training and testing data. The third frame shows the contents of the two Process Documents operators, numbers 3 and 8. The final frame shows the contents of the Cross Validation operator, number 6 in the main frame.

Looking at the initial results achieved on the unseen data, without class balancing, for the Naive Bayes learner the main disappointment is the precision of the model. Although the overall accuracy of 81.44% is good the precision value of 40.57% is low. A similarly disappointing result was achieved when using the K-NN model. By addressing the class imbalance in the unseen test data better results, more in-line with the training results, were achieved.

Results and Discussion

This section will begin by providing the complete set of results achieved whilst developing the model, shown in Table 1, and then discuss.

Before discussing the results from the training models it should be noted that the Accuracy value shown also includes the precision and recall performance for the non bullying class. The initial performance results from the model, whilst not disastrous, were not as strong as would have been expected. However by reviewing the word vector produced it was clear that the tokens being used as input for the learner were not unique enough. By balancing the classes and by replacing the stop word and stemming operators with an N-Gram operator the improvements were seen immediately. Introducing new learners, K-NN and SVM, showed a small improvement in the performance of the model. The final evolution of the model was to introduce feature weighing which had various effects on the model performance. Using feature weighting with Naive Bayes gave the best overall model accuracy at 86% and a very good recall value of 91%. The improvement in performance for K-NN was not as significant and though the overall accuracy and recall for SVM decreased, recall drastically, the precision of the feature weighted SVM model gave the highest results for any test execution. As a final test a feature weighted SVD model was tested and although it accuracy and precision value were poor its recall value was the best seen.

When testing on the unseen data, Table 2, the Naive Bayes model returned the best results. Though not as good as the performance achieved in training the model the results are still impressive.

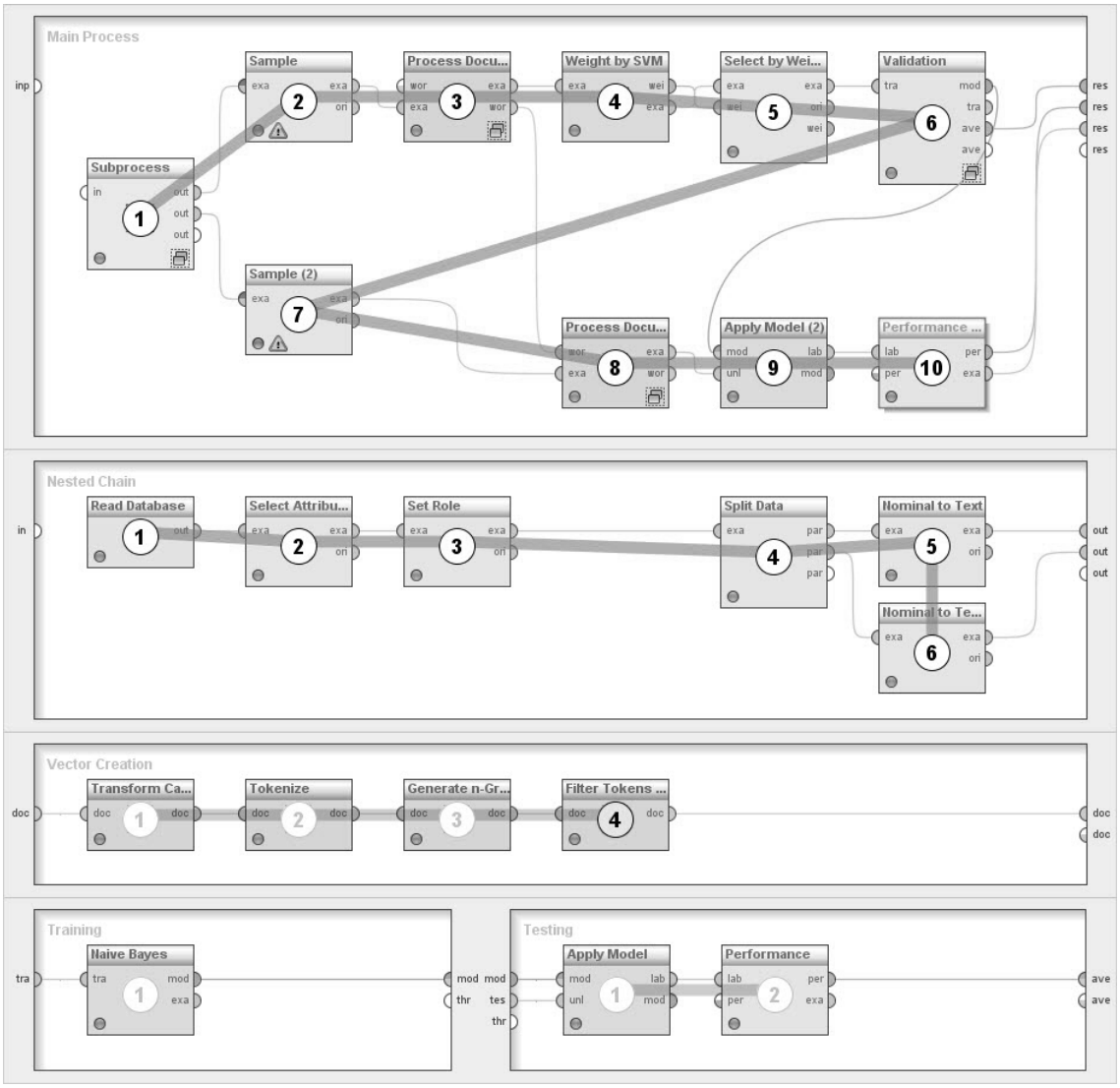


Figure 1: Final Model showing training and testing phases

A sample of results from other cyberbullying detection research in the literature is given in Table 3. Direct comparison between the results from this research and the other research papers presented should be avoided as different datasets, using text in a different environment, are used. It can, however, be seen that the approach used in this research shows great potential, worthy of further effort.

6 Conclusions and Future Work

In this paper the Ask.fm site was introduced and data scraped from the site was used to automatically detect the presence of cyberbullying text. From this work it was seen that nearly one in every seven posts in the sample data could be considered as bullying or abusive. When classifying a question as bullying consideration was given to text that could be perceived as sexually, culturally, physically or mentally abusive. That there is evidence of cyberbullying on the Ask.fm site is not a surprising statement but it must be made. Although a user can apply their own filtering to censor or block continuously abusive users it appears that the site itself does nothing to prevent any form of bullying.

There are some areas where there is scope to continue the work of this project including the scraping of the data, data cleansing including the removal of questions where the language is not English, better utilisation of the scraped data and finally the development of a more realistic model. In real day to day

Table 1: Performance results achieved during model development

Model Description	Accuracy	Precision	Recall
Naive Bayes - Initial Model	48.67%	17.33%	71.81%
Naive Bayes - Balanced Classes	63.94%	60.27%	82.02%
Naive Bayes - N-Grams	72.02%	74.19%	67.66%
K-NN - N-Grams	73.88%	74.29%	73.19%
SVM - N-Grams	77.71%	82.17%	70.96%
Feature Weighted Naive Bayes - N-Grams	86.17%	80.11%	91.16%
Feature Weighted K-NN - N-Grams	77.23%	88.33%	68.09%
Feature Weighted SVM - N-Grams	65.64%	96.23%	32.55%
Feature Weighted SVD - N-Grams	57.90%	54.61%	94.80%

Table 2: Performance results achieved during model Testing

Model Description	Accuracy	Precision	Recall
Naive Bayes	81.44%	40.57%	73.19%
Naive Bayes - Balanced	77.87%	78.85%	76.17%
Naive Bayes - Balanced and Clean	77.40%	78.07%	76.07%

Table 3: Comparison of Model Results with other publications

Paper	Data Source	Accuracy	Precision	Recall
This paper	Ask.fm	78%	79%	76%
Dadvar el al. [3]	YouTube	-	72%	45%
Dinakar et al. [5]	YouTube	63%	-	-
Dadvar el al. [2]	MySpace	-	31%	15%
Yin el al. [12]	MySpace	-	35%	22%

use any model developed would need to be able to be queried with individual questions to determine if they are bullying or not and the model itself would also have to be updated, perhaps in batch mode, where all questions that are determined as non-bullying and questions that are flagged or reported as abusive or bullying are added to the model to improve its accuracy. The development of the model would also have to be better thought out, possibly by supplementing it using natural language processing, as it may not be feasible to add all questions, bullying or not, into the model without it becoming unmanageable.

Other areas where there is scope for future work is to perform sentiment analysis on all of a users given answers. If a user is continuously receiving bullying posts and the sentiment for their answers in becoming more and more negative it should be possible to trigger an alarm that the user may be at risk. It would also be good if certain key words and phrases could be included in the model to identify answers that suggest self-harm or suicidal tendencies.

References

[1] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 71–80.

[2] Dadvar, M., de Jong, F., Ordelman, R., and Trieschnigg, R. (2012). Improved cyberbullying detection using gender information.

[3] Dadvar, M., Trieschnigg, D., Ordelman, R., and Jong, F. (2013). Improving cyberbullying detection with user context. In Serdyukov, P., Braslavski, P., Kuznetsov, S., Kamps, J., Rger, S., Agichtein, E., Segalovich, I., and Yilmaz, E., editors, *Advances in Information Retrieval*, volume 7814 of *Lecture Notes in Computer Science*, pages 693–696. Springer Berlin Heidelberg.

[4] Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3):18:1–18:30.

[5] Dinakar, K., Reichart, R., and Lieberman, H. (2011). Modeling the detection of textual cyberbullying. In *The Social Mobile Web*.

[6] Kontostathis, A., Edwards, L., and Leatherman, A. (2009). Chatcoder: Toward the tracking and categorization of internet predators. In *PROC. TEXT MINING WORKSHOP 2009 HELD IN CONJUNCTION WITH THE NINTH SIAM INTERNATIONAL CONFERENCE ON DATA MINING (SDM 2009). SPARKS, NV. MAY 2009.*

[7] Kontostathis, A., Reynolds, K., Garron, A., and Edwards, L. (2013). Detecting cyberbullying: query terms and techniques. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, pages 195–204, New York, NY, USA. ACM.

[8] Riege, R. (2013). Cyber-bullies claimed lives of five teens. http://www.herald.ie/news/cyberbullies-claimed-lives-of-five-teens-29043544.html. Visited October 14, 2013.

[9] U.S. Department of Health and Human Services (2013). What is bullying, bullying definition. http://www.stopbullying.gov/what-is-bullying/definition/index.html. Visited October 16, 2013.

[10] Vinita Nahar, Xue Li, C. P. (2013). An effective approach for cyberbullying detection. In *Communications in Information Science and Management Engineering*, pages 238–247.

[11] Webwise (2013). Ask.fm: A guide for parents and teachers. http://www.webwise.ie/AskfmGuide.shtm. Visited November 29, 2013.

[12] Yin, D., Davison, B. D., Xue, Z., Hong, L., Kontostathis, A., and Edwards, L. (2009). Detection of harassment on web 2.0. In *PROCEEDINGS OF THE CONTENT ANALYSIS IN THE WEB 2.0 (CAW2.0) WORKSHOP AT WWW2009*.

Measuring and Utilising Diversity in Ensemble Machine Learning Algorithms

Brian Carter[1], Laura Keyes [2]

[1] Department of Informatics, Institute of Technology Blanchardstown, Dublin, Ireland
brianthomascarter@gmail.com

[2] Department of Informatics, Institute of Technology Blanchardstown, Dublin, Ireland
laura.keyes@itb.ie

Abstract

Ensemble learners popularity in machine learning has grown rapidly driven by their positive empirical results and the development of online parallel computer processing. Three seminal ensemble methods, *Bagging, Adaboost and Random Forest* are evaluated and compared on fifteen popular *UCI* datasets. Bagging had the most consistent results, followed by Random Forest. *Diversity*, considered a cornerstone of successful ensemble algorithms has, as of yet, no universal agreed or sufficient measure. Twelve measures were applied to the fifteen datasets. Four measures, *co-incidence failure diversity (CFD), genearlised (GD), pairwise double fault (DF)* and an information theoretic measure, are identified as the most consistently informative across the range of datasets. A greedy backward pruning of the ensemble base classifiers, based on the four identified measures of diversity is deployed to determine if ensemble test set error can be improved. Pruning of the ensembles indicates strong links between the measures of diversity and accuracy, notably with Adaboost. Pruning is a substandard method for improving accuracy compared to increasing the number of base classifiers.

Keywords: AdaBoost, Bagging, Random Forest, Diversity, Pruning

1 Introduction

Ensemble methods of machine learning forego the single model approach and instead generate multiple models, referred to as *base classifiers*. Given the input data x, the data is passed to multiple learners, *base classifiers*, which individually apply a single learning algorithm to the input data, generating a prediction. A *combiner* is then used to aggregate the predictions to form one overall prediction. In order for the combined prediction to be optimal, the individual *base classifiers* or *hypothesis* must be *different* from one another. The difference between base classifiers is referred to as *diversity*. There are many empirical studies showing ensembles outperform single learners (Maclin and Opitz, 2011; Oza and Tumer, 2008). There are four key components required to successfully build powerful ensemble learners.

Diverse base classifiers: In order to generate diverse base classifiers, an *unstable learning algorithm* is preferred. Decision Trees and Neural Networks are unstable learners, in that they are very sensitive to small changes in their parameters. Perturbing the input data x of a Decision Tree can lead to divergent mapping functions $f : \mathcal{X} \mapsto \mathcal{Y}$. Reducing the depth of a Decision Tree makes them more stable with respect to perturbations. Zhou (2012) showed that unpruned trees produce better results in an ensemble classifier. In contrast Support Vector Machines, K-Nearest Neighbour classifiers and Decision Stumps are stable learners, relatively robust to changes in the input data. These algorithms are rarely used as the base classifier in ensembles. This paper limits research to unpruned Decision Trees.

Independence: Independent ensembles generate all the base classifiers in parallel with no communication between base classifiers as is the case in Bagging and Random Forest. Adaboost is dependent.

Combiner: There are many methods for combining the base classifier outputs to form the final hypothesis or prediction. Stacking methods are often used in heterogeneous ensembles, where the base classifiers are not of the same type. In this research, the ensemble models are homogeneous and majority voting combination is used throughout.

Output: Supervised learning with single model algorithms and by extension ensembles can be broken in two types; regression and classification. Classification output can be either actual labels or probability estimates. The focus of this research is on nominal classification labels.

1.1 Research Focus

There is, as of yet, no agreed or sufficient universal measurement of diversity. This work considers three ensemble algorithms and examines their performance across a number of datasets. Measurements of diversity are calculated, aimed at identifying the most informative measure, with respect to model accuracy, across range of datasets. A pruning experiment is conducted to determine if the identified measures of diversity can be used to improve a model's accuracy. The research questions are:

(a) Are any one of the three ensemble algorithms better with respect to reducing the generalisation error over a wide variety of datasets?

(b) Is it possible to identify the most appropriate and widely applicable measure of diversity?

Following on from question (b) there has been research aimed at utilising diversity measures to improve the accuracy of ensembles. Kuncheva (2003) and Li et al. (2012) use the kappa diversity measure with the aim of reducing the size of the ensemble whilst improving the accuracy. The former limits the comparison to three base classifiers at a time, using Pareto Optimality.

Li et al. (2012) extends the kappa measures and includes individual base classifier accuracy as a trade-off bound, recognising that the candidates for removal may be very similar to one another and changes should be targeted to those base learners where the accuracy improvement has the largest potential. Banfield et al. (2005) creates a separate *thinning* set to experiment on the impact to accuracy by removing base classifiers, limiting the research to Random Forest only.

Taking this previous research into account, and inspired by Kuncheva (2003) *overproduce and select* method, the third research question is:

(c) Can the generalisation error be reduced by improving the test set error through implementing a greedy backward pruning strategy (build the whole ensemble and iteratively remove base classifiers) ?

A practical focus of this ongoing research is the use of ensemble techniques in parallel distributed file setting, where diversity rather than accuracy may be utilised as a stopping criteria.

Section 2 sets out the three algorithms, Bagging, AdaBoost and Random Forest and details previous comparisons of the algorithms found in literature. Section 3 details four measurements of diversity, pertinent to the results of the experimentation and details how they are applied to the ensemble algorithms. Section 4 details the methodology of experiments and the results of the questions set out. Finally Section 5 concludes with reflections on the results and areas of possible future research.

2 Ensemble Techniques

Three ensemble algorithms were considered for comparing results across the datsets. The *Bagging Algorithm*, an acronym derived from **B**ootstrap **agg**regat**ing**, is one of the earliest, simplest and still widely used ensemble re-sampling techniques (Re and Valentini, 2011). Given D, a labeled training dataset with m examples over $\mathbf{x}$ predictor variables and y corresponding labels in the format, $\{(\mathbf{x_1}, y_1\}, \ldots (\mathbf{x_m}, y_m)\}$.

For T iterations a subset D^{IB} with m examples is generated by sampling with replacement. A base learner $\mathcal{L}$ is applied at each iteration to D^{IB}. At each iteration 2/3 of the examples are selected. The final combined outputted label is the majority vote after T iterations.

Adaboost is the most widely known boosting algorithm. The base learners are trained sequentially and misclassified examples at each iteration receive higher emphasis, thus increasing their probability of being chosen at following iterations. Each base learner is weighted by its accuracy and the aggregation of the base classifiers is completed with respect to the weightings. Later iterations of the algorithm generate base classifier that tend to be less accurate as they contain more harder to classify and noisy examples.

Random Forest are a hybrid of the Bagging algorithm and the Random Subspace method. The base learner is restricted to unpruned Decision Trees. As in Bagging each tree is constructed with a bootstrapped sample of the dataset D with replacement. The significant difference with Bagging is that at each non terminal branch of the Decision Tree forming the base classifier, a random selection of the available input feature space is taken. The number of attributes selected is typically set to 1/3 of available attributes, but can be varied. Indeed, Bagging can be said to be a special case of a Random Forest where all the attributes are chosen. In Random Forest, the resulting base classifiers are aggregated by majority voting.

2.1 Ensemble comparisons in research

Maclin and Opitz (2011) noted that Bagging is consistently better than boosting across a wider range of datasets, but that boosting can perform much better on certain datasets. In the presence of noise, increasing the number of iterations of the ensemble, will not result in an increase in error for Bagging, but may indeed with boosting, due to overfitting of the noisy examples.

A major drawback of boosting is that each base classifier is dependent on previous base classifier for receiving a weight update for the examples, therefore limiting boosting by the sequential processing power available. Boosting is not easily parallelisable, unlike Bagging.

Random Forest can handle a large number of variables, is easily parallelisable and tolerant to noise. Convergence however depends on the presence of a number of predictor variables (Biau, 2012). In the presence of ambiguity within the data, Random Forest may not perform well. Genuer et al. (2008) observed that in the presence of a large number of input variables, increasing the value of m_{try} (the number of input features to select) and the number of iterations T can improve the generalisation error. Random Forest is more efficient than Bagging and boosting using Decision Trees in terms of computation power.

3 Diversity

Diversity of the base classifiers is fundamental for creating robust ensemble models. Bagging and Random Forest generate diversity implicitly through the effect of random sampling at each iteration. García-Pedrajas et al. (2007) state diversity should be greater in Random Forest over Bagging due to the two random sampling stages. Boosting, by targeting hard to classify examples, explicitly increases diversity, at the expense of accuracy. However the development of ensembles methods did not explicitly incorporate diversity. Rather the emergence of diversity is heuristic and judged by the success of the empirical results.

Kuncheva (2003) sets out to question whether an explicit measure of diversity could be used to improve ensemble models. Diversity measures can be broken into pairwise, non-pairwise and information theoretic measures. Many of the diversity measures use oracle outputs, as shown in Table 1.

Consider that in any training set, there will be easy and hard to classify examples. A paradox or trade-off exists between accuracy and diversity. As the individual base classifier increases accuracy, it is not feasible to continually increase diversity (Stapenhurst, 2012). As accuracy increases, diversity reduces as the base classifiers will have more agreement on correct predictions. From Table 1 the maximum

	D_j correct (1)	D_j wrong (0)
D_i correct (1)	a	c
D_i wrong (0)	b	d

Table 1 Oracle output relationship between two classifiers

diversity would be achieved if all the correct classifications on D_i were incorrect on D_j. There should be independence between the errors of the base classifiers.

3.1 Diversity Measurements

The *Double Fault* measure, (Giacinto and Roli, 2001) is very intuitive. It is a measure of the proportion of times two classifiers both mislabel the same object. A higher value indicates reduced diversity as there are more coincident errors.

$$DF_{i,j} = d \Rightarrow D_{DF} = \frac{1}{L(L-1)} \sum_{i=1}^{L} \sum_{k \neq i}^{L} \frac{d_{i,j}}{N} \tag{1}$$

Where d is the count of times both comparable classifiers are wrong, as in Table 1 and L the number of base classifiers.

Generalised diversity is based on the idea that diversity is maximal when an incorrect labeling in one classifier is accompanied by the correct labelling in another. (Partridge and Krzanowski, 1997). The probability of k classifiers correctly predicting a randomly chosen example is $P(c_i = \frac{k}{L})$. Given this, the generalised diversity over all pairs can be calculated as in equation (2). The range of the output is $[0, 1]$ with a higher value indicating greater diversity.

$$D_{GD} = 1 - \frac{\sum_{i=1}^{L} l(l-1)P(c_i = \frac{L-l}{L})}{\sum_{i=1}^{L} l(L-1)P(c_i = \frac{L-l}{L})} \tag{2}$$

Concident failure diversity modifies D_{GD} to ensure that the minimum value is 0 when all the classifiers are correct or when they are all are simultaneously correct or wrong on the same examples (Partridge and Krzanowski, 1997). The maximum value 1, is achieved when all the mis-classifications are unique.

$$D_{CFD} = \begin{cases} 0 & \text{if } Pr(c_i = 1) \\ \frac{\sum_{i=1}^{L}(L-l)Pr(c_i = \frac{L-l}{l})}{(L-1)(1-Pr(c_i=1))} & \text{otherwise} \end{cases} \tag{3}$$

Brown (2009) attempted to extend the restrictive nature of pairwise measures of diversity by developing an *Information theoretic* based diversity. This comprised of three terms.

Relevancy is the measure of mutual information between each of the outputs of the base classifiers $X_{1:T}$ and the true label Y. A large value increases diversity.

Redundancy is the measure of the interaction between the possible subsets of the set of classifiers S. A large value indicates that there are strong correlations among the classifiers and intuitively reduces diversity. This value should be minimized.

Conditional redundancy is the strength of correlation between all possible subsets conditioned on the class label. While the goal is to strive for low correlation among the classifiers, high correlations are preferred when the base classifiers predict the class label correctly.

The calculation of the correlations on all the possible subsets of base classifiers S is very complicated and there is in fact no effective means for calculating these when the number of base classifiers is only moderately large (Zhou and Li, 2010). Therefore Brown resolves to only find the pairwise measures as an approximation.

Index	Set	No. Class	Exam -ples	Attri- butes	Index	Set	No. Class	Exam -ples	Attri- butes
1	ionosphere	2	351	34	9	breast	2	699	9
2	sonar	2	208	60	10	diabetes2*	2	768	8
3	iris	3	150	4	11	soybean	19	307**	35
4	wine	3	178	13	12	heart	2	270	13
5	vechicle	4	846	18	13	creditg	2	1000	20
6	glass	6	214	9	14	hepatitis	2	155	19
7	segmentation	7	2310	19	15	credita	2	690	15
8		2	768	8					

*The diabetes dataset is replicated in diabetes2. The Pima Indian diabetes dataset contains records across 8 features. 5 of the features contain values that are biologically impossible, e.g. zero value for body mass index etc. The diabetes2 has these values removed. ** The soybean dataset has 19 classes. 1 class *"2-4-d injury"* has only one entry and was removed.

Table 2 Datasets used in ensemble algorithm/diveristy measure analysis

$$I_{brown}(X_{1:T};Y) \approx \underset{\text{diversity}}{\sum_{i=1}^{T} I(X_i;Y)} \quad \underset{\text{relevancy}}{} \quad \underset{\text{redundancy}}{-\sum_{i=1}^{T}\sum_{j=i+1} I(X_i,X_j)} \quad \underset{\text{conditional redundancy}}{\sum_{i=1}^{T}\sum_{j=i+1} I(X_i,X_j|Y)} \tag{4}$$

Having taken note of the difficulty in calculating Brown's diversity measure, Zhou and Li (2010), put forward an alternative method for estimating the diversity within the limitations of pairwise comparison.

They postulate that rather than taking all possible joint distributions when calculating the redundancy and conditional redundancy it can be reduced to (5), where S is a set = $\{X_1, \cdots X_{i-1}\}$ and S_k is a subset.

$$I_{zhou}(X_{1:T};Y) \approx \underset{\text{diversity}}{\sum_{i=1}^{T} I(X_i;Y)} \quad \underset{\text{relevancy}}{} \quad \underset{\text{redundancy}}{-\sum_{i=1}^{T} max_{S_k} I(X_i;\{S_k\})} \quad \underset{\text{conditional redundancy}}{+\sum_{i=1}^{T} max_{S_k} I(X_i;\{S_k\}|Y)} \tag{5}$$

4 Experimental Work

Fifteen well known datasets, used widely in other research, were selected from the *UCI Machine Learning Repository*. They were selected to include a full spectrum of characteristics; all continuous variables, all discrete variables, mixture of both; binary and multiple classes; presence of noise. The datasets are listed in Table 2. A brief description of the information contained in each dataset can be found at `http://archive.ics.uci.edu/ml/`.

The ensembles were built with unpruned Decision Trees as the base classifiers and combined by majority vote. The MATLAB programming environment was use to created the ensembles, calculate the measures of diversity and perform pruning experiments.

4.1 Ensemble Algorithm Results

Results obtained in this study, indicate Bagging is predominately the best performing method returning the smallest average classification error on thirteen of the fifteen datasets. AdaBoost performs better on two of the datasets. Random Forest's error is very close to that of Bagging and comes second to Bagging on ten datatsets.

The optimal number of iterations for all three methods was five hundred. This is significantly influenced by Random Forest as it favours a larger number of iterations than Bagging and AdaBoost. Indeed at higher number of iterations Bagging appears to cause the error to deteriorate. Fifty iterations is the overall optimal number of iterations for Bagging, contrary to Maclin and Opitz (2011) .

AdaBoost under-performs the other two algorithms on a number of occasions. AdaBoost is particularly sensitive to noisy data found in the *vechicle, soybean* datasets and performs poorly. AdaBoost also

	↓	↓	↑	↓	↑	↓	↑	↓	↑	↑	↓	↓
	Pairwise				**Non-pairwise**						**Info. Theoretic**	
	Q-stat.	Corre-lation	Dissag-reement	Double Fault	Kohavi Wolpert	Iter-ator	Ent ropy	Diff-culty	Gen. Diff.	Co-incid failure	Brown*	Zhou*
% All	0.567	0.389	0.3	0.767	0.3	0.433	0.244	0.689	0.711	0.778	0.778	0.759
% Bag	0.633	0.4	0.267	0.833	0.267	0.533	0.2	0.833	0.733	0.867	0.889	0.889
% Boost	0.6	0.5	0.433	0.667	0.433	0.5	0.367	0.533	0.7	0.633	0.667	0.611
% RF	0.467	0.267	0.2	0.8	0.2	0.267	0.167	0.7	0.7	0.833	0.778	0.778

* Results given for 9 binary class datasets. Arrow indicate whether an increase/decrease of the value indicates more/less diversity.

Table 3 Tracking diversity measure on error, assumption greater diversity $\approx$ reduced error

performs poorly where there are multiple classes, as in the **glass** dataset, where the classes are not evenly distributed.

Random Forest performs poorly on *creditg* dataset. The *creditg* contains seven continuous attributes and thirteen nominal attributes. Recoding the nominal attributes into dummy variables significantly reduces the error, on a par with Bagging and AdaBoost. The increase in randomization, by introducing a greater number of input variables has a significant impact. This is not the subject of this research but poses an interesting research question regarding the data pre-processing steps required to optimise Random Forest's performance values.

4.2 Diversity Measures Analysis and Results

From the results of the analysis of the ensemble algorithms detailed in Section 4.1 the analysis of diversity measures was to limited to 25 and 50 iterations of each ensemble. The information theoretic measures are calculated on actual predicted labels, not oracle outputs, and are thus are limited in this research to datasets with binary class output (9 datasets).

The twelve measures of diversity for each ensemble algorithm for both variances of number of iterations were calculated. A simple sloping heuristic was applied to determine if any one of the measures consistently indicated increased diversity in the presence of reduced test set error. The results of the percentage of positively correlated examples are presented in Table 3.

Overall, across all datasets, algorithms and iterations, co-incidence failure diversity(CFD) is the best tracker of the link between diversity and accuracy, followed by double fault(DF), generalised diversity(GD) and the information theoretic measures. Section 3.1 presents a discussion of these measures. For a good treatment of the other measures of diversity see Kuncheva (2003).

The recognition of the paradox/trade-off between accuracy and diversity identified by Stapenhurst (2012) is borne out in the three non-information theoretic measures, in that diversity measurement should focus on mis-classified examples.

4.3 Over-produce and Select

The contribution of each base classifier to the overall diversity of the ensemble, using the four identified measures was calculated. The base classifiers were iteratively removed, with the base classifier contributing the least to overall diversity being removed first. This strategy was limited to Bagging and AdaBoost for twenty-five iterations only.

The results indicate that while there is no consistent improvement across the datasets, there are individual improvements. AdaBoost achieves better improvements than Bagging, achieving on average a 10% improvement when D_{GD} and D_{CFD} diversity measures are used to target the base classifiers. However in contrast the same methodology increases the Bagging test set error on average by 21%.

It should be noted, that the improvements in accuracy, tend to be seen on those datasets that had a low test set error, that is were easy to model. In general there would not appear to be significant advantage,

taking computational complexity into account, in choosing a pruning strategy over simply increasing the number of base classifiers to an optimal amount.

5 Conclusions

The paper presented an empirical comparison of three ensemble algorithms, reviewed and evaluated measures of ensemble diversity and investigated whether these measures could be used to improve the algorithms. A number of corroborative conclusions with previous research were found.

Maclin and Opitz (2011) compared Bagging and Adaboost using pruned Decision Trees and Neural Networks. Many of the datasets they used are similar to this work. A comparison with those results, indicates that Bagging using unpruned Decision Trees and a higher number iterations of the ensemble learner, as in this study, outperforms Bagging as applied by Maclin and Opitz.

A strong link between the error rates of the individual base classifiers of an AdaBoost ensemble and their contribution to the overall diversity of the ensemble was observed. This is in line with Brown et al. (2005) differentiation between *explicit* diversity creation using boosting methods and *implicit* diversity creation using bagging methods.

With respect to the four optimal diversity measures identified, Kapp et al. (2007)'s investigations, focusing solely on Adaboost came to similar conclusions for two of the measures, generalised(GD) and double fault(DF). Stapenhurst (2012) found similar evidence to support the optimality of the double fault measure. Gangadhara and Dubbaka (2010) found that double fault and coincidence failure diversity(CFD) were more strongly correlated with ensemble error in the context of random subspace sampling.

This study also found both information theoretic measures of diversity returned similar values, confirming that Zhou and Li (2010) meausure of diversity, using maximal subset, does approximate Brown (2009) measure.

5.1 Future Research

The literature review revealed conflicting claims regarding the effects of using pruned and unpruned Decision Trees and varying the number of base classifiers. A strict and rigorous study using well known datasets could provide clarity.

Random Forest performed much better using dummy binary variables in the presence of discrete attributes with a large number of levels. Two knowledge gaps are identified; Can this observation be exploited to identify the interactions between various levels in separate discrete input variables and is it possible to identify the type of data that would benefit from such pre-processing in order to optimize the results using Random Forest?

The research into the improvement of the ensemble error, through pruning using diversity measures, indicated that in a minority of cases it was possible to affect a reduced error. Additionally a wider investigation may determine if there are underlying conditions that allow this to be the case for some datasets over others.

The datasets under investigation in this study are well known and relatively small in size. Further research into combining measures of diversity with ensemble learning in a distributed parallel environment with large volumes of data could be considered. This may yield benefits in the context of uneven or unknown data distributions.

The measures of diversity have a high computational complexity. The benefits of calculating the diversity measures versus simply increasing the number of iterations of an ensemble must be considered. Along this line of thought it is prudent to conclude with some concerns as aired by Ho (2002).

Rather than guiding research efforts towards finding the best subset of input features and the best applicable classifier, there is now a move towards seeking the best set of classifiers and the best combination method. Very soon Ho states, the search will move towards the best set of combination methods and the best way to use them all, and ensemble research will enter a infinite loop driven by every more

complicated combination schemes, theories and ad-hoc approaches and lose sight of the true strength and finesse of the original solution.

References

Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. (2005). Ensemble diversity measures and their application to thinning. *Information Fusion*, 6(1):49–62.

Biau, G. (2012). Analysis of a random forests model. *The Journal of Machine Learning Research*, 98888:1063–1095.

Brown, G. (2009). An information theoretic perspective on multiple classifier systems. In *Multiple Classifier Systems*, pages 344–353. Springer.

Brown, G., Wyatt, J., Harris, R., and Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion*, 6(1):5–20.

Gangadhara, K. and Dubbaka, S. A. R. (2010). Comparing compound diversity and ordinary diversity measures using decision trees. Master's thesis, University of Boras.

García-Pedrajas, N., García-Osorio, C., and Fyfe, C. (2007). Nonlinear boosting projections for ensemble construction. *The Journal of Machine Learning Research*, 8:1–33.

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008). Random forests: some methodological insights. *arXiv preprint arXiv:0811.3619*.

Giacinto, G. and Roli, F. (2001). An approach to the automatic design of multiple classifier systems. *Pattern recognition letters*, 22(1):25–33.

Ho, T. K. (2002). Multiple classifier combination: Lessons and next steps. *Series in Machine Perception and Artificial Intelligence*, 47:171–198.

Kapp, M. N., Sabourin, R., and Maupin, P. (2007). An empirical study on diversity measures and margin theory for ensembles of classifiers. In *Information Fusion, 2007 10th International Conference on*, pages 1–8. IEEE.

Kuncheva, L. I. (2003). That elusive diversity in classifier ensembles. In *Pattern Recognition and Image Analysis*, pages 1126–1138. Springer.

Li, N., Yu, Y., and Zhou, Z.-H. (2012). Diversity regularized ensemble pruning. In *Machine Learning and Knowledge Discovery in Databases*, pages 330–345. Springer.

Maclin, R. and Opitz, D. (2011). Popular ensemble methods: An empirical study. *arXiv preprint arXiv:1106.0257*.

Oza, N. C. and Tumer, K. (2008). Classifier ensembles: Select real-world applications. *Information Fusion*, 9(1):4–20.

Partridge, D. and Krzanowski, W. (1997). Software diversity: practical statistics for its measurement and exploitation. *Information and software technology*, 39(10):707–717.

Re, M. and Valentini, G. (2011). Ensemble methods: a review.

Stapenhurst, R. (2012). *DIVERSITY, MARGINS AND NON-STATIONARY LEARNING*. PhD thesis, the University of Manchester.

Zhou, Z. (2012). *Ensemble Methods: Foundations and Algorithms*. Chapman & Hall.

Zhou, Z.-H. and Li, N. (2010). Multi-information ensemble diversity. In *Multiple Classifier Systems*, pages 134–144. Springer.

Comparison of Video Game Classification Schemes Utilising Web Mining Techniques

Sheamus Causer [1], Markus Hofmann [2]

[1] Institute of Technology Blanchardstown, Dublin, Ireland
sheamus.causer@gmail.com

[2] Institute of Technology Blanchardstown, Dublin, Ireland
markus.hofmann@itb.ie

Abstract

During the period 2002-2012, the video games industry more than doubled in size confirming its' evolution from niche interest to a multi-billion dollar entertainment industry with mass appeal. However, the growing demand for games targeted at mature audiences and the increasing visual realism of games, has resulted in calls across a number of jurisdictions for a tightening of controls around children's' access to potentially harmful content. In response, a number of industry and government championed classification schemes have emerged around the globe, all of which share a common goal of seeking to ensuring children only have access to games which are age appropriate. However, differing social and political attitudes towards childhood development and how criteria relating to violent, sexual and other content should be applied, has resulted in inconsistencies in minimum age requirements between geographical regions. This issue is further compounded by the ability to purchase games online (thereby bypassing local jurisdiction assessments) and the lack of awareness parents/guardians may have about the criteria underpinning different classification schemes. The following paper utilises a variety web mining techniques to compare the classifications awarded to games released in Europe and the US.

Keywords: Web Mining, classification

1 Introduction

Fuelled by the rise of mobile gaming, the advent of 'blockbuster' titles and the launch of next generation home gaming consoles, the global sales of video games in 2012 exceeded USD66bn with consumer spending expected to reach USD78bn by 2017 (Nayak, 2013). The growing visual realism of games and incorporation of adult themes has resulted in a number of jurisdictions implementing classification schemes similar to those applied in the movie industry. These classification schemes seek to provide consumers (and retailers) with guidance as to the age appropriateness of a game based on the level of violence, coarse language and other selected criteria.

As with any form of censorship, determining what is 'appropriate' and for whom is a complex social and political issue, resulting in a variety of schemes being used in different countries. For example, in Australia prior to the introduction of the 'Restricted (R18+)' category in January 2013, video games depicting high levels of violence or sexual themes were refused classification and hence, were effectively banned for sale. As a result, a range of internationally released titles (including 'Left for Dead 2', 'Grand Theft Auto IV' and 'Fallout 3') were either banned or modified to achieve a lower classification. Interestingly one of the key arguments put forward by the proponents of the 'Restricted (R18+)' category, was that the majority of gamers in Australia were in fact over 18 years old (with the average age now approaching 30), both dispelling the 'teenager' stereotype associated with gaming, and emphasizing the growing commercial importance of content targeted at mature audiences (Delfabbro and King, 2010). This shift in play demographics is also reflected in the US where females over the age of 18 (31%) represent a significantly greater proportion of players than boys aged under 18 (19%) (esa, 2013).

The Australian debate also highlighted the challenges of trying to define and enforce censorship classifications in an interconnected world. The increasing proliferation of gaming enabled devices,

coupled with the ability to download software from other jurisdictions enables children to bypass many of the physical controls (e.g. identification checking at retail stores) traditionally used to bar access to a product. Even when online access is done under adult supervision, variations between jurisdictions in terminology, acceptable age thresholds and classification criteria add extra layers of complexity. To further explore this challenge, the following paper examines how the application of two different schemes (PEGI and ERSB) can affect the classification of games released in European and US markets.

2 Video Game Classification Process

In the context of the video game industry, classification schemes seek to prevent individuals from being exposed to material that they are likely to perceive as unsettling or harmful. Whilst this is generally framed in the context of protecting children (and hence the close association between classification classes and audience age), these schemes also seek to prevent the distribution of material that is deemed incompatible with what Wilson (2008) describes as 'the public good'. The banning in Australia in June 2013 of the original version of 'State of Decay' (not due to its use of high impact violence but because of the game's inclusion of sequences deemed to glamourize the use of illegal narcotics), highlights the at times subjective and contradictory nature of classifying games for older audiences (and the fine line between classification and censorship).

Whilst variations exist between jurisdictions, typically the process for obtaining a game classification follows a similar path. Prior to being made available for sale, games are submitted by publishers to the classification authority with jurisdiction over that region, which will then evaluate and assign a rating based on published criteria. (For globally released titles this will require the game to be evaluated by multiple authorities which can lead to variations in the classification of the same game between regions.) Accompanying a copy of the game will also be a 'content declaration' completed by the publisher providing a detailed walk through of the game and in some instances a self-evaluation of what classification should be awarded. As the publisher has the most insight into the content of the game, the content declaration provides a means for classifiers to identify and focus on that content which is most likely to affect the overall classification (pegi.info).

As the classification awarded may restrict sales to certain age groups (potentially shrinking the size of the customer market), publishers may in some instances actively seek to achieve a lower classification to improve the potential commercial return. Hence, it is not unusual for publishers to appeal an assessment, and where necessary customize the content to achieve the desired classification. Conversely for games targeted specifically at older audiences, achieving a restricted classification may carry with it an important cachet which gives the game more credibility/desirability within that genre.

Depending upon the region, restrictions on marketing may also be applied prior to the game being officially classified (hence, games that have been reviewed on games websites prior to release may be flagged as 'not yet classified'). Restrictions may also be applied following classification to what screenshots and other marketing material can be placed in the public domain for more restricted classifications. In addition, to advertising the classification, a number of rating schemes also require publishers to include key words (e.g. 'mild violence', 'coarse language', etc.) to provide context as to why the rating was assigned.

2.1 Classification Schemes

As the primarily focus of video game classification is the protection of children, classification classes are generally differentiated based on the minimum age of the audience for whom the content is deemed appropriate. However, varying cultural and scientific perceptions on child cognitive development have resulted in a range of classification schemes being developed globally including the 'Pan European Game Information' (PEGI) and 'Entertainment Software Rating Board' (ESRB) schemes (Table 1).

2.1.1 Pan European Game Information (PEGI)

Originally published in 2003, the PEGI scheme was introduced in Europe to assist with streamlining classification schema between countries and is administered by the 'Interactive Software Federation of Europe', a not for profit organisation based in Belgium. PEGI has since been adopted by over 30

European and non-European countries as either their official classification standard (i.e. attaining a PEGI rating is mandatory in order to be able to sell games in that jurisdiction), or integrated with existing national classification process (i.e. whilst PEGI rated products can be sold the Irish Film Censorship Office retains the right to reclassify games based on their content). As part of the PEGI code of practice, publishers are also required to ensure the classification awarded is consistent with the requirements of local legislation if selling into other markets.

Table1: Video Game Classification Schemes

Entertainment Software Rating Board (ESRB)	Pan European Game Information (PEGI)
Early Childhood (eC) Content is intended for young children.	
Everyone (E) Content is generally suitable for all ages. May contain minimal cartoon, fantasy or mild violence and/or infrequent use of mild language.	**3** Suitable for all ages. May contain mild violence in an appropriate context for younger children, but no explicit language is allowed.
Everyone 10+ (E10+) Content is generally suitable for ages 10 and up. May contain more cartoon, fantasy or mild violence, mild language and/or minimal suggestive themes.	**7** Suitable for ages 7 and older. May contain mild, cartoon violence, sports, or elements that can be frightening to younger children.
Teen (T) Content is generally suitable for ages 13 and up. May contain violence, suggestive themes, crude humour, minimal blood, simulated gambling and/or infrequent use of strong language.	**12** Suitable for ages 12 and older. May contain violence in a fantasy setting, coarse language, mild sexual references or innuendo, or gambling.
Mature (M) Content is generally suitable for ages 17 and up. May contain intense violence, blood and gore, sexual content and/or strong language.	**16** Suitable for ages 16 and older. May contain explicit violence, strong language, sexual references or content, gambling, or drug use (encouragement).
Adults Only (AO) Content suitable only for adults ages 18 and up. May include prolonged scenes of intense violence, graphic sexual content and/or gambling with real currency.	**18** Suitable for ages 18 and older. May contain graphic violence, including "violence towards defenseless people" and "multiple, motiveless killing", strong language, strong sexual content, gambling, drug use (glamorization), or discrimination.

As at December 2012, a total of 20,305 games had undergone PEGI classification since its inception as shown in Table 2 (PEGI, 2012). For the year 2012, a total of 1,820 games were submitted for rating, of which 170 (9.4%) of games were awarded an '18' classification.

Table 2: Breakup of number of games by PEGI classification level [source: pegi.info (2014)]

Classification	Number of games	%
3	9,350	46.0%
7	2,761	13.6%
12	4,467	22.0%
16	2,502	12.4%
18	1,225	6.0%
Total	**20,305**	**100.0%**

2.1.2 Entertainment Software Rating Board (ESRB)

Used predominantly by retailers in the US and Canada, ESRB operates on a similar model to PEGI (scheme administered by a not for profit organisation with publishers contractually agreeing to abide by a code of practice). Whilst the categories used by the two schemes are broadly comparable, variations in minimum age limits and criteria for levels of acceptable sexual, violence and gambling content reflect differing social norms between the regions. For example whilst the inclusion of 'explicit violence' (reaches a stage that looks the same as would be expected in real life) is permitted

for audiences as young as 16 years under PEGI (vs. 17 years under ESRB), the number of games that have been further restricted to audiences over 18 year due to violent content under ESRB is considerably less (7 vs. 100+ for PEGI). Comparisons of the number of games by classification level, indicates significantly less occurrences of Adults Only being awarded to titles under ESRB (Table 3).

Table 3: Breakup of number of games by ESRB classification level (source: esrb.org)

Classification	Number of games	%
Early Childhood	289	0.90%
Everyone	21,070	65.86%
Everyone 10+	2,419	7.56%
Teen	6,223	19.45%
Mature	1,960	6.13%
Adults Only	32	0.10%
Total	**31,993**	**100.0%**

2.2 Classification datasets

Details of what classification has been assigned to a game can be obtained from a variety of sources. In addition to classifications being printed on packaging and marketing materials, both ESRB and PEGI schemes operate websites where consumers can search by game title. The popularity of video gaming has also led to the inclusion of game reviews on a number of media, technology and dedicated games websites, which in some instances may include the game's classification.

To determine how a child's access to potentially harmful games' content may vary between regions, the ESRB and PEGI classifications for a selection of 128 console (PS3 and Xbox) and PC based games were obtained and compared. To provide an opportunity to explore variations in web-mining techniques required to extract classification data from primary and third party websites, PEGI data was extracted directly from pegi.info whereas ESRB data was obtained from the games website g4tv.com.

3 Web content mining methodology

After reviewing a variety of websites that post video gamer reviews, a US based site was selected as the source for ESRB data as it regularly posts reviews of newly released titles incorporating details of their classification. Whilst games are often ported (converted) to run on a variety of platforms to maximise their commercial success, their content (and hence classification) typically remains the same. Hence, games sites tend to review titles only once on a single platform (with links pointing back to this review) unless the release on another platform is considered significant.

Utilising the 'Crawl Web' operator in RapidMiner ('RM'), an individual webpage for each game review was extracted. As the directory path for reviews on the site was dependent upon which platform the review was done on (and review pages contain multiple tabs for linked videos, blogs, etc.), three tailored versions of the crawl website process were developed to enable extraction of reviews performed on Xbox, PS3 and PC.

The resulting 128 reviews obtained were then processed to extract selected attributes including the game's title ('game') and classification rating ('censor') using XPATH based attribute extraction techniques (two games were identified as having multiple reviews, 'NHL 13' and 'Tony Hawks Proving Ground', for their respective releases on Xbox and PS3). At this stage, pages linked to games still awaiting censor classification (censor = 'rating pending') were excluded from further analysis. At the completion of processing, the resulting dataset was exported into MS Excel where the games' titles form the basis of the approach used to obtain their corresponding PEGI rating.

3.1 Obtaining PEGI classification dataset

Using the pegi.info game search facility, the title of one of the games was selected from the MS Excel dataset and manually entered to identify the URL to be utilised to search for the corresponding PEGI ratings. The naming convention used by the search query is identifiable in the resulting URL, and provides a template for converting the MS Excel dataset into a list of URLs for further web crawling.

As the game title is embedded in the game title, the URL is copied into MS Excel and split into three segments to enable a separate URL to be tailored for each other game in the dataset.

- Segment 1: http://www.pegi.info/en/index/global_id/505/?searchString=
- Segment 2: <game title>
- Segment 3: &agecategories=&genre=&organisations=&platforms=& countries=&submit=Search#searchresults

The column containing game titles ('game') is copied from the ESRB list and the spaces between words in the title are replaced with '+' to enable use in the search string (i.e. 'Binary Domain' becomes 'Binary+Domain'). The segments of the URL are rejoined following insertion of the game title, with the resulting list forming the pages to be crawled to obtain the corresponding PEGI data.

Utilising RM, the pegi.info URL list is loaded utilising the 'Read Excel' operator which then forms the target list for the 'Get Pages' operator. The returned pages are then passed through a 'Process Documents' operator which acts as a wrapper for a series of steps including stripping the html tags from the pages and tokenizing the resulting contents. Although classifications on pegi.info appear in the form of images (.gif), the naming conventions used for the image files can be used to identify what level has been awarded (e.g. icon18.gif = 18). Tokenizing by linguistic expression enables the preservation of the text string that identifies which .gif appears on the page which can then be identified in the resulting word list.

As the only attributes being sort are those that identify the game and its' associated classification, the 'Filter Tokens (by Content)' operator is applied to exclude terms that do not include the string './img/ratXS/' which is used to identify the classification label (filtering is also further applied to remove attributes not ending in '.gif'). The resulting dataset is exported to MS Excel where use of nested if/then statements converts the multiple attributes into a single column with the PEGI classification. Utilising the URL used to obtain the PEGI dataset as a primary key, and the VLOOKUP function within MS Excel, the PEGI and ESRB classification datasets are then combined to form a single dataset which forms the basis for further analysis.

3.2 Mapping classification datasets

Out of the 128 reviews obtained from the review website, 37 games (28%) were unable to be mapped to pegi.info using this methodology. Manual searching for these games on pegi.info identified that a number of these games had been submitted for classification but due to variations in how their titles were written (e.g. inclusion of a sequel's extended title, use of punctuation in titles, etc.) the associated URL had not been returned by the 'Get Pages' operator. For example, the full title for "Risen 2: Dark Waters" is used on the review website (and also by esrb.org), however the title is shortened to "Risen 2" on pegi.info. (Hence, for future research an iterative approach using multiple variations of game titles may be considered to improve sample recall).

4 Comparison of ESRB and PEGI classification assignment

The resulting video games dataset incorporates 91 titles (that have both PEGI and ESRB classifications) spread across a range of genres. Note: the 37 titles that were not matched between the two source datasets were excluded from further analysis. To provide a basis for comparison and analysis, the classifications used by the two schemes were converted into a single 5 point scale based on the matrix shown in Table 4 (with 1 = least harmful content and 5 = most harmful content).

Table 4: Standardized 5 Point Classification Scale

		Entertainment Software Rating Board (ESRB)		Pan European Game Information (PEGI)	
		Description	Minimum Age	Description	Minimum Age
Scale	1	(eC) / (E)	All Ages	(3)	All Ages
	2	(E10+)	10	(7)	7
	3	(T)	13	(12)	12
	4	(M)	17	(16)	16
	5	(AO)	18	(18)	18

On this basis, 58 titles (59.8%) were identified as has having inconsistent classifications between the two schemes (an extract of which is provided in Table 5) with six of these titles having variations of 2 or more points. Classifications under the PEGI scheme (average: 3.52) were generally higher than those assigned by ESRB (average: 3.07).

Table 5: Games (including ESRB and PEGI 5 point scale classification)

Game Details		Classification		ESRB (5 point scale)					PEGI (5 point scale)				
Genre	Title	ESRB	PEGI	1	2	3	4	5	1	2	3	4	5
Action	Alan Wake's American Nightmare	T	18										
Action	Asura's Wrath	T	16										
Action	Borderlands 2	M	18										
Action	Call of Duty: Black Ops 2	M	18										
Action	Dragon's Dogma	M	18										
Action	I Am Alive	M	18										
Action	Kingdoms of Amalur: Reckoning	M	18										
Action	Mass Effect 3	M	18										
Action	Max Payne 3	M	18										
Action	NeverDead	M	18										
Action	Ninja Gaiden 3	M	18										
Action	Prototype 2	M	18										
Action	Sleeping Dogs	M	18										
Action	Starhawk	T	16										
Action	Syndicate	M	12										
Action	TERA	M	12										
Action	Twisted Metal	M	12										
Action	Yakuza: Dead Souls	M	18										
Adventure	Jet Set Radio	T	12										
Adventure	Sorcery	E10+	12										
Fighting	Dead or Alive 5	M	16										
Fighting	Skullgirls	T	16										
Fighting	Tekken Tag Tournament 2	T	16										
Fighting	UFC Undisputed 3	T	16										
Fighting	WWE '12	T	16										
Platform	Deadlight	M	18										
Racing	Little Big Planet Karting	E	7										
Racing	Trials Evolution	E10+	12										
RPG	Final Fantasy XIII-2	T	16										
RPG	Game of Thrones	M	12										
Shooter	Battleship	T	16										
Shooter	Binary Domain	M	18										
Shooter	Dishonored	M	18										
Shooter	Far Cry 3	M	18										
Shooter	Hitman: Absolution	M	18										
Shooter	Inversion	M	18										
Shooter	Spec Ops: The Line	M	18										
Simulation	NHL 13	E10+	12										
Simulation	NHL 13	E10+	12										
Sports	MLB 2K11	E	12										
Sports	Tony Hawk's American Wasteland	T	16										
Strategy	Journey	E	7										
Strategy	Sins Of A Solar Empire: Rebellion	T	7										
Strategy	XCOM: Enemy Unknown	M	18										
Survival-Horror	Resident Evil 6	M	18										

4.1 Analysis of variations

Whilst the largest source of variation was due to games receiving lower classification levels under ESRB than PEGI, instances of the opposite were also identified. Perhaps unsurprisingly the majority of variations centred around genres traditionally associated with depictions of violence and use of high level adult themes (e.g. Action / Shooter / Survival Horror). Interestingly, variations were also identified among sporting (e.g. MLB 2k11) and racing games (e.g. Little Big Planet) due to the use of mild coarse language and 'non-realistic looking violence' towards fantasy characters. Arguably however, differences of up to one point for a game (when using the standardised 5 point scale) can be attributed to the different minimum age limits used by the two schemes (which may result in games 'just' missing a threshold).

4.1.1 Access to adult content

Whilst the level of harm posed by one point variations at lower classification schemes is debatable, both schemes recognise the concept of adulthood occurring at 18 years (level 5) and the requirement to restrict such content from younger audiences (which in some jurisdictions may also trigger additional restrictions over the promotion and distribution of such games). Hence, the assignment of games to either levels 4 or 5 acts as a 'de facto' binominal classifier for assessing the availability of adult content between the two schemes (and a one point variance becomes significant). Hence the high level of variation shown between PEGI[5]:ESRB[4] in Table 5 is significant as it highlights not only differences in perceptions of what should be restricted to adults only, but potential variances in the legal restrictions in place over the same material in different jurisdictions. On this basis, adult content (predominantly in the form of games containing higher levels of violence) appears more accessible under ESRB.

In addition to differences over the use of Level 5 classification, another means of identifying other potential key points of differentiation between the two schemes are games which varied by greater than one point utilising the standardized rating scale (and hence have more than 'just' missed a cut off age limit). Four games were identified as having variations of 2 points including 'Alan Wake's American Nightmare', which highlights the differences in weighting assigned to the inclusion of 'violence against defenseless people' under the different schemes.

.4.1.3 Variations between genres

As per books, movies and other entertainment media, the allocation of games to genres provides audiences with a means of determining if the subject matter (e.g. 'Puzzle', 'Action', etc.) or style of play (e.g. 'Shooter' typically indicates 1st person) matches their interests (with classification then assisting with determining if the content is age appropriate). Hence, a number of genres due to the inherent nature of their content will attract higher classification ratings (i.e. 'Survival Horror') than others (i.e. 'Puzzles'), a pattern which is reflected by both schemes. However, a high number of variations arose due to lower ESRB ratings for games linked to genres typically associated with violence (action [18], shooter [7], fighting [4]), indicating generally higher levels of tolerance for violent content under ESRB. This is further reinforced by the absence of games being rated as only being suitable for adults.

5 Conclusion

The ongoing growth of the video gaming industry and its adoption by an increasing older audience highlights the importance of classification schemes to assist with preventing minors from being exposed to content deemed inappropriate and/or harmful. The lack of a single global entity with sufficient authority to define and enforce classification levels, has resulted in the development and adoption of varying schemes in different jurisdictions (including ESRB in US/Canada and PEGI in Europe). Whilst the number and nature of classification levels used by these schemes is comparable, varying social and academic attitudes towards the age appropriateness of violence, coarse language and other content can result in the same game being available to significantly different audiences depending on the scheme used. An issue that is further exacerbated by the growing trend towards online distribution, which enables minors to bypass many of the traditional physical controls used to prevent access to restricted content. As classification can also have a financial impact by narrowing a

game's target audience, this borderless marketplace coupled with competing classification schemes, creates a potential risk of 'shopping' between schemes to obtain the desired rating. All of which highlights the importance of parents/guardians being aware of the differences in the criteria that underpin the classification labels used by these schemes.

Bibliography

Delfabbro, P and King, D., 2010, "Should Australia have an R 18+ classification for video games?", Youth Studies Australia, Volume 29, pp: 9-15

Entertainment Software Association (esa), 2013, "Sales, demographic and usage data", pp:1-16

Nayak, M., 2013, "A look at the $66 billion video-games industry", Reuters

"PEGI Annual Report – 2012", Interactive Software Federation of Europe (ISFE)

www.pegi.info/en/index/

Wilson, S., 2008, "Censorship, new technology and libraries", The Electronic Library, Volume 26, Issue 5, pp: 695-701

www.esrb.org/index-js.jsp

g4tv.com

www.classification.gov.au

Session 2

Data Analysis

BIG DATA OR SMALL DATA? – A TELECOMMUNICATIONS SCENARIO

Eloy Martinez Colomina [1,2], Sheila Fallon [2] , Enda Fallon [2], Yuansong Qiao [2]

[1] Ericsson Software Campus, Athlone, Ireland
eloy.martinez@ericsson.com

[2] Software Research Institute, Athlone Institute of Technology, Ireland
emartinez@research.ait.ie, sheilafallon@ait.ie, efallon@ait.ie, ysqiao@research.ait.ie

Abstract

Big Data systems have recently emerged for Internet level data processing. Telecommunications systems have traditionally been thought of as generating large scale data; performance logs, alarms, counters. This work investigates the applicability of Big Data systems to telecommunications data processing. In particular this paper investigates how the structure and size of the data produced by a telecommunications network affects the performance of the Hadoop file system. The Hadoop file system is a framework based on the map reduce model. It is designed to process large data sets (Big Data). This paper investigates the performance of Hadoop for processing telecommunication data. Results presented illustrate that restructuring the telecommunications data into a more Internet like format is shown to achieve significant performance improvement.

Keywords: 3GPP, Big Data, Map Reduce, XML.

1 Introduction

The emergence of major Internet companies, e.g. Google, Amazon, and Facebook, has resulted in challenges dealing with huge quantities of data. NoSQL databases are designed for dealing with this large scale data. While Telecommunications systems have traditionally been thought of as generating large scale data, the scale of this data is somewhat small in comparison to the volume produced by Internet companies. In addition content produced by Internet companies such as Twitter is unstructured in nature whereas telecommunications data, which is specified by standards such as 3GPP, is semi structured in format. This paper investigates how the volume and characteristics of telecommunications data impacts on the performance of the Hadoop File system.

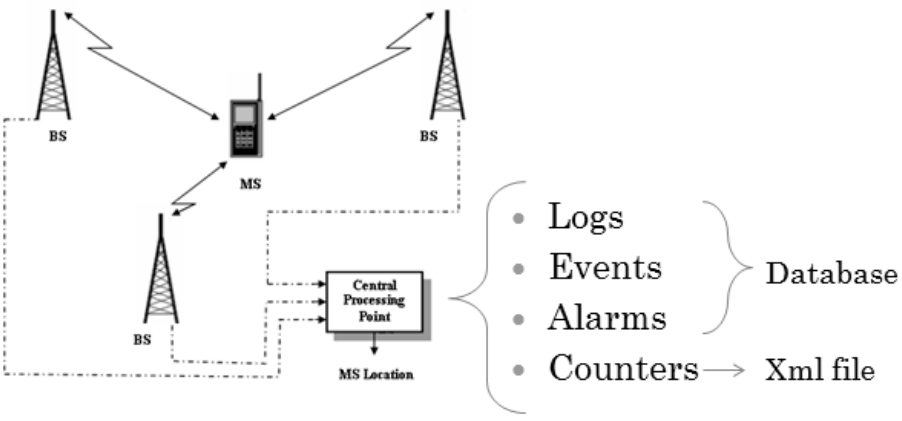


Diagram 1: Data Produced by Base Stations

The evaluation considers how (a) input format (b) data structure (c) file size affects the performance of Hadoop's map reduce algorithm when processing Xml files. Results illustrate that the characteristics of a telecommunications environment; a large number of small files and 3GPP structure have a significantly detrimental effect on performance. Restructuring the telecommunications data into bigger file size is shown to achieve significant performance improvement.

This paper is structured as follows. Section 2 outlines related work in the area. An overview of relevant technologies is provided in Section 3. Results are presented in Section 4. Finally conclusions and future work is outlined in section 5.

2 Related Work

A number of papers have considered the implications of small file input on Hadoop file performance. In [2], the authors describe and design a new architecture for small file storage using Hadoop. The focus of the paper is store small files into incomplete blocks. The load of name node is reduced using the DataNode to store metadata. The approach outlined in [2] uses unstructured small files, the approach outlined in this paper how the semi structured nature of 3GPP performance data effects performance. In [3] the authors focus on PowerPoint (PPT) files, merging small files and using prefetching to mitigate the load of the NameNode. The result is a large file made off the pictures of the PPT and an index with information about the length of the PPTs. In such an approach the index and some prefetching records are needed to maintain the system. The approach outlined in this work does not require prefetching as all necessary information is encapsulated in the file. In our approach the merging stage is achieved with minimal load.

[4] adds a small file processing module in internal HDFS. The small file processing module creates a large file within a special structure (index, length and content) and two running process: merge and read. For the approach outlined in this paper the files should be processed together and already have structure. Such structure provides an advantage as the additional structure and process outlined in [4] are not required. [5] modifies the HDFS to reduce the metadata footprint in the NameNode's main memory. The approach alters the block structure by extending the start of the block with an offset and length of the sequenced files. The semi structured nature of 3GPP performance data again provides a performance advantage in comparison to such an approach.

A number of papers provide a general overview of Hadoop storage of telecommunications data. The [7] and [8] analyses the potential for Hadoop to process and analyze the large data set generated by service providers in the Chinese market. In [9] a Hadoop-based network traffic analysis system to perform high-speed data-intensive network traffic analyses is proposed. [10] utilizes Hadoop to detect performance limitations regarding the relationships between hosts. [1,7-10] illustrate the potential of Hadoop to process telecoms data in order to extract contextual meaning.

3 Technical Overview

3.1 3GPP (Performance IRP)

The 3rd Generation Partnership Project (3GPP) unites 6 telecommunications standard development organizations (ARIB, ATIS, CCSA, ETSI, TTA, TTC), known as "Organizational Partners" and provides their members with a stable environment to produce the highly successful Reports and Specifications that define 3GPP technologies. [11]

The original scope of 3GPP was to produce globally applicable Technical Specifications and Technical Reports for a 3rd Generation Mobile System based on evolved GSM core networks and the radio access technologies that they support (i.e., Universal Terrestrial Radio Access (UTRA) both Frequency Division Duplex (FDD) and Time Division Duplex (TDD) modes). The scope was subsequently amended to include the maintenance and development of the Global System for Mobile

communication (GSM) Technical Specifications and Technical Reports including evolved radio access technologies. [11]

The 3GPP caters to a large majority of the telecommunications networks in the world. It is the standard body behind UMTS (Universal Mobile Telecommunications System), which is the 3G upgrade of GSM. Most cellular networks on the planet are based on GSM.

3.2 Hadoop (Map Reduce)

Hadoop is an open source framework that implements the MapReduce parallel programing model in commodity hardware. Hadoop includes two parts: HDSF and Map Reduce. [12]

The Hadoop distributed file system (HDFS) is designed to store and process big files. The files stored are split into blocks. The size of the block is equal for the whole cluster but can be configured (32, 64, 128… MB). HDFS blocks are large compared to disk blocks. The reason is to minimize the cost of seeks, by making a block large enough, the time to transfer the data from the disk can be significantly longer than the time to seek to the start of the block. Thus the time to transfer a large file made of multiple blocks operates at the disk transfer rate. The HDFS stores small files inefficiently, since each file is stored in a block, and block metadata is held in memory by the NameNode. However, small files do not take up any more disk space than is required to store the raw contents of the file. For example, a 1 MB file stored with a block size of 128 MB uses 1 MB of disk space, not 128 MB [13].
As a result Hadoop suffers a performance penalty when storing a large amount of small files.

The map reduce model consist of 2 parts: map and reduce. The map phase is the part of the implementation that runs over the data and extracts the content. When dealing with small files the NameNode suffers a performance penalty especially in terms of memory. In such a situation the Jobtracker is obligated to generate an excessive number of maps thereby affecting the Map phase. The time to process a small file in the map phase is less than a large file. However the number of maps reduces efficiency.

Figure1 represents a generic Hadoop cluster.

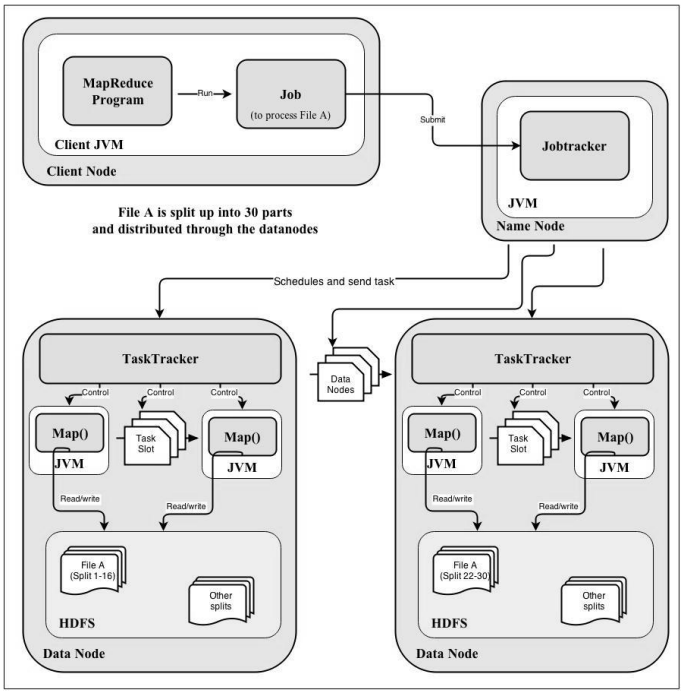


Figure 1: Generic Hadoop cluster

4 Results section

4.1 Experimental Environment

The experimental configuration consisted of a Cloudera 4.5 Hadoop cluster. The cluster uses 2 machines an Intel Xenon E5540 (2.53 GHz CPU, 64-bit, 16 cores) with 110GB of memory and a 880 GB hard disk. Each node had Red Hat 6.3 OS installed and Java version 1.7. The block size was 128 MB and the number of replications was 3. The nodes were interconnected by a 1.0 Gbps Ethernet network.

4.2 An Analysis of Input format

This section evaluates how the format used by Hadoop to input telecommunications performance data affects map reduce processing time. 2 input formats are evaluated:

- *"XmlInputFormat"* - Within Hadoop (Mahout Project [14]) this class is specifically intended for XML file processing. Using this input format the size of the file is not relevant. When parsing the input file the XmlInputFormat will only consider data contained between tags defined by the user.

- *"CombineFileInputFormat"* - This abstract class have been extended to work with XML files in the same way as XmlInputFormat, extracting data contained between tags.

The main difference between the XmlInputFormat and CombineFileInputFormat input format types is the number of maps created per class. The XmlInputFormat creates a single map per file. If the input file is bigger than the block size the XmlInputFormat creates the same number of maps as blocks in the input file. This configuration results in 2 possible scenarios (a) XmlInputFormat processing small files (less than the block size) (b) XmlInputFormat processing large files (greater than the block size).

Alternatively "CombineFileInputFormat decouples the amount of data that a mapper consumes from the block size of the files in the HDFS."[13**]**. In this scenario less mappers are generated to process the same amount of information as a larger volume of data is processed by one mapper (when processing small files).

These input formats are the best choices due to the *XmlInputFormat* is specific to work with xml files (out type of data) and *CombineFileInputFormat* is the class provided by Hadoop that try to avoid the problem with the small files. The rest of input formats provided by Hadoop are not even close to be chosen as candidate because are designed for others purposes.

Figure 2 illustrates the processing time for various quantities of data. The data is presented for evaluation in 2 formats (a) small input file format (as current generated by a LTE mobile system) (b) as a single large file consisting of an amalgamation of the smaller input files (time to merge the files is negligible). Both of the input file formats XmlInputFormat and CombineFileInputFormat are evaluated.

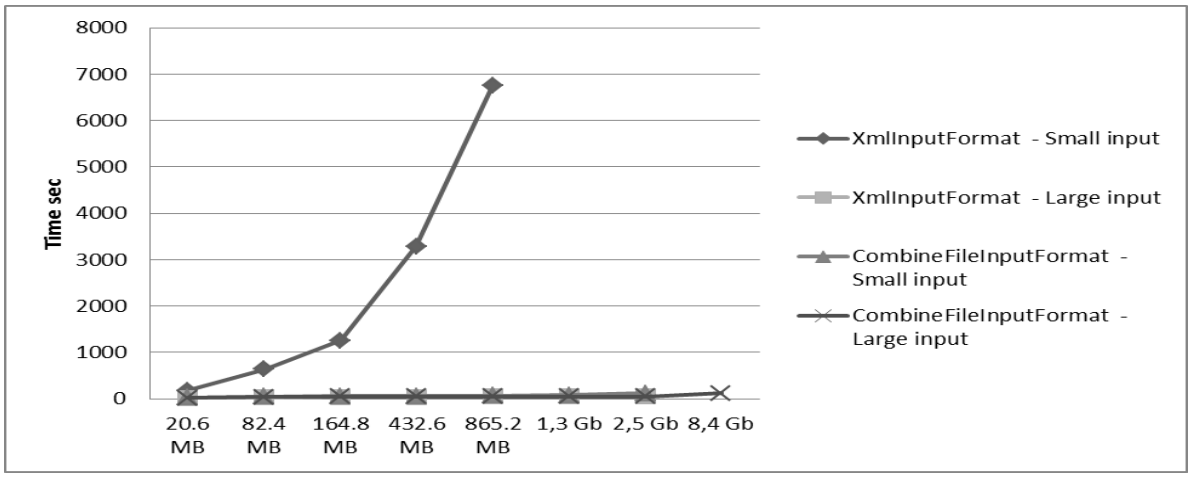


Figure 2: Processing time of XmlInputFormat and CombineFileInputFormat

Figure 2 illustrates significant performance degradation when processing small files using XmlInputFormat when the volume of data increases. When processing 865MB of data in small files the processing time is almost 7000 seconds. The source of the performance degradation is the creation of a map to process each file. These maps process only a very small amount of data and the time to launch and configure them on the HDFS is larger than the time of process the data. As a result the parallelization of the maps is almost impossible. In our configuration with 32 map slots the number of maps produced was 574 for the smallest size folder. However, using CombineFileInputFormat this degradation is avoided as the generation of maps does not overflow.

Table 1 details the number of maps per configuration. It illustrates that using XmlInputFormat with either a single large file or using the CombineFileInputFormat the performance degradation is avoided.

Size Folder No. Small files	**20 MB 574**	**80 MB 2296**	**164 MB 4592**	**432 MB 12054**	**865 MB 24108**	**1,3 GB 36162**	**2,5 GB 72324**	**8,4 GB 241080**
Xml Small	574	2296	4592	12054	24108	-	-	-
Xml Large	1	1	2	4	7	11	21	68
Combine Small	1	1	2	4	7	11	21	-
Combine Large	1	1	2	4	7	11	21	68

Table 1: Maps per job

Figure 3 provides a more fine grained view of the data from Figure 2. Figure 3 illustrates that although CombineFileInputFormat is faster than XmlInputFormat with small files, the use of a large file instead of small files has a significant positive effect on processing time.

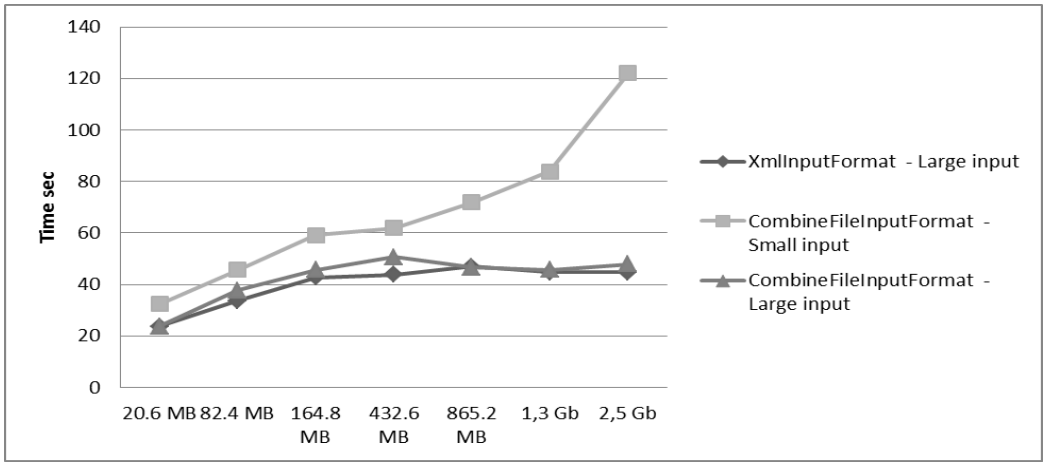


Figure 3: Processing time deleting XmlInputFormat small files input

Despite CombineFileInputFormat has a good performance with small files table 1 shows that has a limitation with the number of small files due to memory issues.

4.3 An Analysis of Data Structure

In our configuration the structure of the XML files adhered to 3GPP version 6.2.0 TS 32.401. The structure of the XML file is as follows:

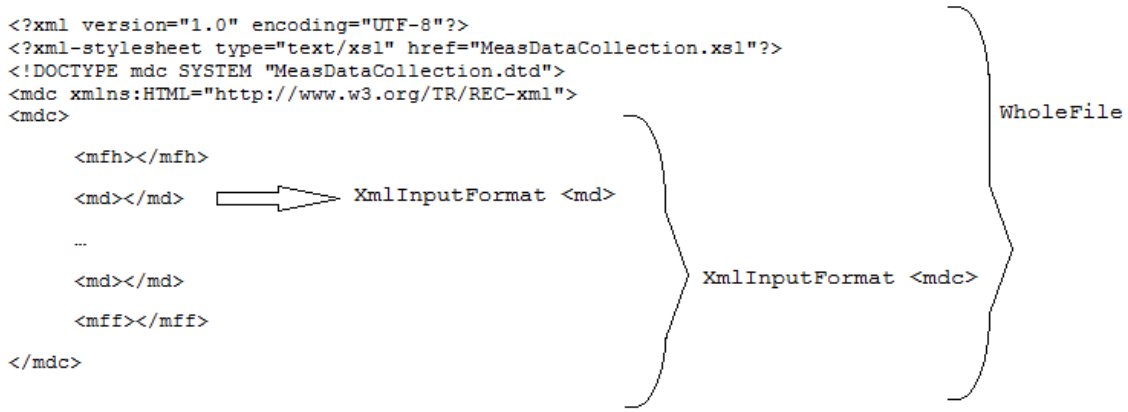


Figure 4: Basic structure XML.

In this section the structure of the XML file is used to change the input to the mapper. Figure 4 illustrates that the file can be considered as 3 sub structures; whole file, mdc partition and md partition.

The whole file represents all content within the XML file. The XML file can be split up in 2 parts: Header and <mdc>. The Header is equivalent in all XML files, whatever the source or version. The Header protocol however does not provide any relevant performance data. Moreover it is a minimal portion of the whole and can be discarded during parsing. In effect the content bounded by the <mdc> tags represents the relevant performance data in the file.

The <mdc> section is subdivided in 3 different parts: <mfh>, <md> and <mff>.

MeasFileHeader	mfh	fileHeader	This is the measurement result file header to be inserted in each file. It includes a version indicator, the name, type and vendor name of the sending network node, and a time stamp ("collectionBeginTime").
MeasData	md	measData	The "measData" construct represents the sequence of zero or more measurement result items contained in the file. It can be empty in case no measurement data can be provided. The individual "measData" elements can appear in any order. Each "measData" element contains the name of the NE ("nEId") and the list of measurement results pertaining to that NE ("measInfo").
MeasFileFooter	mff	fileFooter	The measurement result file footer to be inserted in each file. It includes a time stamp, which refers to the end of the overall measurement collection interval that is covered by the collected measurement results being stored in this file.

Table 2 : Extract of 3GPP TS 32.401

The <mfh> and <mff> tags contain little relevant data and are disregarded. Any relevant data from the <mfh> and <mff> sections is replicated elsewhere in the file. The data relevant to the map reduce process is contained within the <md> tags.

Figure 5 illustrates the use of XmlInputFormat to process 5 large files of varying size. The XmlInputFormat is used to parse both the <md> and <mdc> structures.

Figure 5 illustrates improved performance when parsing the <mdc> rather than <md> tags. The variation in performance is not as a result of the maps since the input file and the block size are equal. The variation in performance results from the number of map input records. Map input records are the number of input records consumed by all the maps in the job. They are incremented every time a record is read from a RecordReader and passed to the map's map() method by the framework[13].

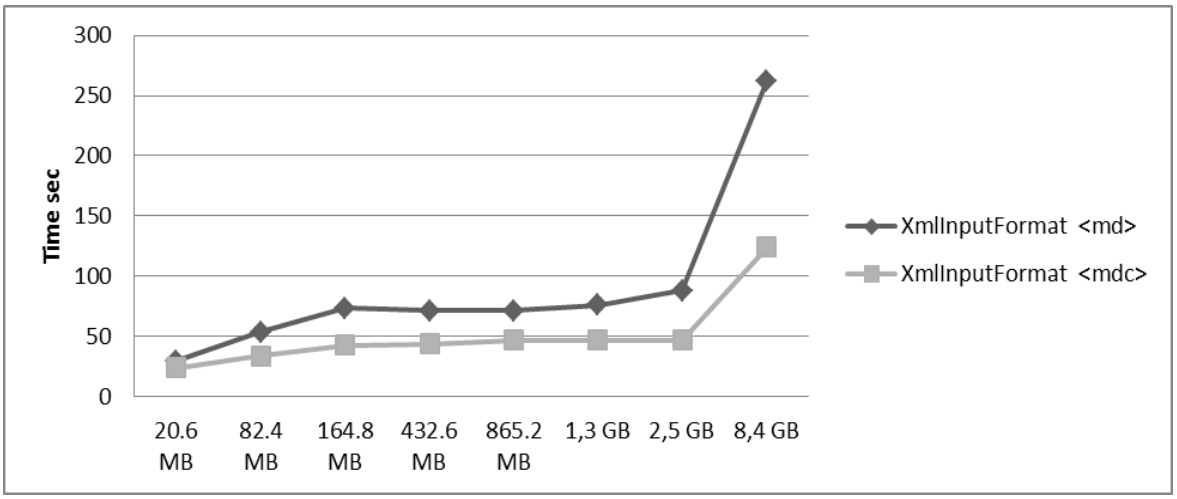


Figure 5: Processing Time of XmlInputFormat <md> versus XmlInputFormat <mdc> Structure

Each file contains one main <mdc> structure. However inside each file there exist an undetermined number of <md> tags. When the XmlInputFormat is used to parse the <mdc> tags the input document is split up fewer times than when parsing <md> tags. The reduction of the number of splits reduces the time required to process the file.

Folder Size No. Small files	**20 MB 574**	**80 MB 2296**	**164 MB 4592**	**432 MB 12054**	**865 MB 24108**	**1,3 GB 36162**	**2,5 GB 72324**	**8,4 GB 241080**
Xml MDC	574	2296	4592	12054	24108	36162	72324	241080
Xml MD	10906	43624	87248	229026	458052	687078	1374156	4580520

Table 3 : Records per job

Table 3 illustrates the total number of records for each job. The <mdc> pattern has the same amount of records as the amalgamation of small files is made of, for example for a folder of 20MB size there are 574 small files and 574 <mdc>elements. In our configuration there are 19 <md> elements within each <mdc> element, for example for a folder of 20MB size there are 574 <mdc> elements and 10906 <md> elements.

Figure 5 illustrates a change at point 164.8 MB in the slope. This change is because at this point the size of the files is bigger than the block size and the parallelization start to work. The change observed between 2.5 GB and 8.4 GB is normal, due to the high increase of data (3 times bigger).

For 4.2 and 4.4 sections CombineFileInputFormat uses <mdc> structure.

4.4 An Analysis of File Size

This section studies the behavior of Hadoop when the size of the source file is modified. This section only considers the XmlInputFormat parsing <mdc> tags and CombineFileInputFormat as results in previous section showed these parsing mechanisms to be optimal.

Figure 6 illustrates the processing time of a fixed 5GB data input when the size of the input file was altered.

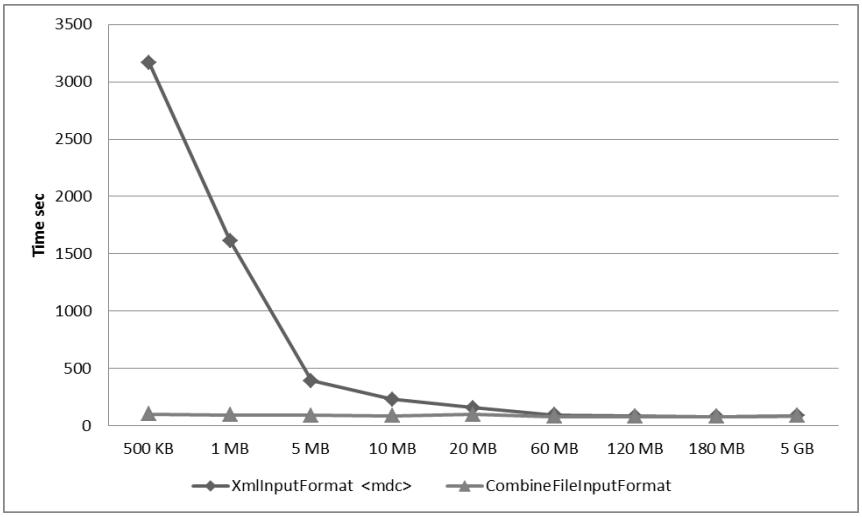


Figure 6: Processing time of 5 GB data input when the size of the file is increased.

As file size is increased the execution time of the map phase reduces sharply for XmlInputFormat, as a result of reducing map numbers. For CombineFileInputFormat however the performance improvement is not as significant as the generation of maps is managed by the split size, not by the number of files. The execution time of XmlInputFormat and CombineFileInputFormat are similar as the file size increases. With a 5GB input file XmlInputFormat has a processing time of 86.813 sec while CombineFileInputFormat has a processing time of 88.796 seconds. Table 4 illustrates the number of maps per job as the size of the input file increases for both XmlInputFormat and CombineFileInputFormat. When 180 MB block is reached we see a strange behavior due to the file is bigger than the block size, so the generation of maps is bigger than the division of the total size between the block size (5120/128=40 approx.).

Size Files No. Files	**500 KB 10250**	**1 MB 5125**	**5 MB 1024**	**10 MB 512**	**20 MB 256**	**60 MB 86**	**120 MB 43**	**180 MB 30**	**5 GB 1**
Xml<mdc>	10250	5125	1024	512	256	86	43	58	40
Combine<mdc>	39	39	38	37	37	29	22	30	40

Table 4 : Maps per Job

5 Conclusion and Future Work

The emergence of major Internet companies has resulted in challenges dealing with huge quantities of data. NoSQL approaches such as Hadoop are designed to process data of such a scale. While telecommunications systems have traditionally been thought of as generating large scale data, the scale of this data is somewhat small in comparison to the volume produced by Internet companies. This paper investigates how the volume and characteristics of telecommunications performance data impacts on the performance of the Hadoop file system. Results presented illustrate that when utilizing Big Data mechanisms there are significant issues with the manner in which telecommunications data is produced. Typically telecommunications performance data is generated at intervals as a number of small files. Results presented in this paper illustrate that the processing time of those files (large number) creates a bottleneck in the NameNode and an intolerable creation of maps to process a relative small amount of data.

In order to address the performance degradation file input format was analyzed. Results presented illustrate that merging small files into a single large file reduces processing time from 6980 seconds to 48 seconds. Results also illustrate the limitation of merging files before processing. To avoid the performance limitation it is suggested to utilize the CombineFileInputFormat parsing mechanism.

An analysis of data structure was undertaken in which two mechanisms to extract content were investigated; <md> partition and <mdc> partition. Results illustrate the time to process the <md> structure is 1.67 times greater than the <mdc> partition for file sizes 164.8 MB to 2.5 GB. For file sizes of 8.4 GB the processing time was 2.11 times greater. The variation in performance increases as more files are added.

Finally an analysis of file size is undertaken which analyses the performance of merged file size. Merged file size is an important consideration as greater file size improves performance. The challenge is to determine at what file size performance is acceptable. Results illustrate a reduction in processing time when we increase the input file size. However performance isn't comparable to large file processing until the block size is reached since distribution of tasks is optimized based on block size [13].

Future work will analyze the implications of random text creation on Hadoop processing performance.

6 References

[1] Rob van den Dam. "Big Data a sure thing for Telecommunications". 2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery.

[2] Liu Jiang, Bing Li, Meina Song. "The optimization of HDFS based on small files". Proceeding of IC-BNMT2010.

[3] Bo Dong, Jie Qiu, Qinghua Zheng, Xiao Zhong, Jingwei Li1, Ying Li. "A Novel Approach to Improving the Efficiency of Storing and Accessing Small Files on Hadoop: a Case Study by PowerPoint Files". IEEE International Conference on Computer Science and Service Computing, 2010.

[4] Yang Zhang, Dan Liu. "Improving the Efficiency of Storing for Small Files in HDFS" 2012 International Conference on Computer Science and Service System.

[5] Chandrasekar S, Dakshinamurthy R, Seshakumar P G, Prabavathy B, Chitra Babu. "A Novel Indexing Scheme for Efficient Handling of Small Files in Hadoop Distributed File System". 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 04 - 06, 2013, Coimbatore, INDIA.

[6] Fan Tongke. "Hadoop-Based Data Analysis System Architecture for Telecom Enterprises". 2013 International Conference on Computational and Information Sciences.

[7] DONG Chao, PENG Dong, HUANG Bolun, LEI Zhenming, YANG Jie. "Large-Scale traffic characterization of Chinese multimedia messaging service using Hadoop". China Communications ,September 2013. Communications China, 2013.

[8] Tingting Li, Jun Liu, Zhenming Lei, Yun Xie. "Characterizing Service Providers Traffic of Mobile Internet Services in Cellular Data Network". Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2013.

[9] Jie Yang, Shuo Zhang, Xinyu Zhang, Jun Liu, Gang Cheng. "Analysis of Smartphone Traffic with MapReduce". Wireless and Optical Communication Conference (WOCC), 2013.

[10] Jerome Francois, Shaonan Wang, Walter Bronzi, Radu State, Thomas Engel. "BotCloud: detecting Botnets Using MapReduce". WIFS'2011, Brazil.

[11] "3GPP official site", http://www.3gpp.org/about-3gpp, 2014.

[12] "Hadoop official site", http://hadoop.apache.org/, 2014.

[13] Tom White, Hadoop: The Definitive Guide, 3rd ed. O'Reilly, May. 20012. Pages 45, 77, 240, 261, 31.

[14] "Mahout Project", https://mahout.apache.org/, 2014.

Performance Evaluation and Optimization of A Hybrid Temporal Pattern Mining Algorithm

Jie Deng [1], Zhiguo Qu [2], Yongxu Zhu [1], Gabriel-Miro Muntean[2] and Xiaojun Wang [2]

Dublin City University, Dublin, Ireland
[1] {jie.deng3,zhu.zhuyonz2}@mail.dcu.ie
[2] {zhiguo.qu,gabriel.muntean,xiaojun.wang}@dcu.ie

Abstract

As the amount of data acquired is increasing rapidly driven by the exceptional development of the information and communication technology area, many researchers have looked at data mining solutions for extracting valuable knowledge from this data. One of the most powerful tools in data mining is sequential pattern mining which aims to find sequential patterns in the data. However, most of the sequential pattern mining algorithms focus on point based events, and process one event attribute only and therefore are unsuitable to handle interval based events such as stock market tradings or patient records. In this paper, the performance of a hybrid temporal pattern mining (HTPM) algorithm is evoluated by comparing it to a classic point based algorithm. Furthermore, some suggestions toward improving this algorithm are made which could lead to more patterns found or less mining time. The improved algorithm also can offer users more options when performing data mining and are faced with hardware limitations.

Keywords: Hybrid Temporal Pattern Mining; Sequential Pattern Mining; Data Mining

1 Introduction

Sequential pattern mining is known for processing data and discovering valuable sequential patterns in the dataset. One of the most common example is the association rule mining in retail market: knowing what things customers usually buy together could enable the retailers to adjust their policies and eventually increase their sales [Chu and Lin, 2005]. Another example is personalised web navigation based on the analysis of website access traces: finding which web pages users prefer or are suitable to their device could increase the quality of browsing experience [Muntean C. H., 2001].

The developments of the last decades have enabled sequential pattern mining algorithms to handle a wide diversity of data types, data structures and data dimensions [Han et al., 2007]. A well-known algorithm in this category called PrefixSpan [Pei et al., 2001] uses prefix and project methods to find frequent patterns without searching the database repeatedly. This algorithm is very effective, but does not support complex data types.

Like PrefixSpan, most of the other algorithms are designed to process point based events and very few of them can be used for processing interval based events. This is easy to understand because most of the daily events can be transformed to point based events by adding the time stamp [Han and Kamber, 2006]. However, there is a need to analyse the complex relationship between temporal events,and therefore an algorithm which can process interval events or hybrid events is required. For example, each patient record can be seen as an interval-based event, as each disease the patient suffers from naturally evolves during a specific period of time. An analysis of the patient database from the temporal perspective will help diagnose faster and eventualy cure the disease in the future.

Hybrid temporal pattern identification has many possible applications [Fan et al., 2011]. Considering a telecommunication system for example, the events happening on devices are point based, and data

flow transmission is interval based [Kambayashi et al., 2000]. These two kinds of events should be handled together for analysis. A data mining process can be used to predict what is going to happen in the telecommunication network and consequently give the manager suggestions to avoid system failure [Xiaodan, 2011]. Furthermore, changing of method for generating pattern candidates gives the user more options in terms of reducing the mining time and increasing the number of patterns discovered.
This paper evaluates an algorithm called Hybrid Temporal Pattern Mining (HTPM) [Wu and Chen, 2009]. This algorithm uses a candidate generator and a test to find hybrid event patterns. All the candidates are generated by combining two existing patterns which can reduce a certain work.
In this paper, we firstly discuss some related concepts about sequential pattern mining in section 2. Secondly, an implementation of the algorithm HTPM is specifically elaborated in section 3 including potential improvement. Furthermore, the experiments and results are shown in section 4 which compared the performance of HTPM and PrefixSpan, and the improvements as well.

2 Related Works

Three categories of sequential pattern mining based on event type can be identified: point-based, interval-based and hybrid-based [Wu and Chen, 2007]. The approaches to different problems are quite different. In this section, the mining problems and the main methods used to solve these problems are discussed.

2.1 Point-based events

The most common pattern in sequential pattern mining is frequency pattern, which translates into the number of occurrences. In general this number should be greater than certain threshold, a minimum value, for consideration [Lakshmi and Raghunandhan, 2011].

2.1.1 Generalized Sequential Pattern (GSP)

Instead of growing all the patterns found in the dataset, the GSP algorithm [Hirate and Yamana, 2006] generates only those patterns already proven frequent. This algorithm is using one of the apriori properties: all frequent itemsets should have a frequent parent itemset, respectively. The GSP algorithm can reduce the complexity, but it does not perform well enough when processing lengthy pattern sets.

2.1.2 Sequential PAttern Discovery using Equivalence classes (SPADE)

Unlike GSP which uses a structure like $< sequence\_ID : sequence\_of\_itemsets >$, the SPADE algorithm [Zaki, 2001] uses a vertical id-list database format such as $< itemset : (sequence\_ID, event\_ID)) >$, to associate each sequence a list of its occurances. Then, frequent sequences can be found efficiently using these id-lists. This method also reduces the database scanning times, and further reduces the execution time.

2.1.3 Prefix-projected Sequential Pattern mining (PrefixSpan)

Apriori was the first algorithm introduced in sequential pattern mining. The process of Apriori is simply build each pontential itemsets by existing items, and varify in dataset. Unlike Apriori bottom-up process, PrefixSpan algorithm loads the whole database into the memory first, and finds the frequent items by projecting sub-databases of different prefixes. The project process is to find the first position in a transaction and remove the items before. This algorithm is very effective, but does not support complex data types.

2.2 Interval-based events

Temporal pattern mining discovers interval-based event patterns instead of point-based event patterns in sequential pattern mining. The most complicated part of interval based pattern mining is the relationship between patterns, as there are 13 different positions two intervals could be relative to each other [Villafane et al., 1999]. This is compaletely different from other sequential pattern mining algorithms based on three types of relationship between patterns. Furthermore, two parameters (start time and end time) of each event are considered in temporal pattern mining, that is twice the load of the point event based sequential pattern mining algorithms.

2.2.1 Kam and Fu's method (KF)

Just like the process in Apriori and GSP, the KF method [Kam and Fu, 2000] generates candidates and tests if the candidate is a pattern. The main difference between KF method and Apriori is that the candidates are more complicated for temporal reasons. Just like Apriori in sequential pattern mining, the KF method needs large time and storage to perform data mining.

2.2.2 TPrefixSpan

TPrefixspan is a PrefixSpan-based approach for processing interval-based event[Hu et al., 2009]. Unlike PrefixSpan, each pattern found in the current projected database is attached to the prefix to become a new pattern after the TPrefixSpan handles more complicated circumstances due to the different temporal relationships between patterns.

3 Design and Implementation

The process of mining hybrid patterns is different from that of point-based event algorithms for several reasons: first, the interval-based event cannot be divided simply into two point-based events without a pair-wise relationship guarantee [Mannila, 1996]. Secondly, the ability of processing two occurrence time is required. So the concept of simply separate start time and end time of a interval event is not appropriate.

3.1 HTPM algorithm

As already mentioned above, the basic concept of temporal pattern mining is finding relations between events, and see if the relationship is frequent [Nisbet et al., 2009]. With this concept in mind, it seems intuitive that the simplest solution is a process of finding each event in the database, and using an Apriori-like algorithm to generate all the candidates and test them. However, this is proven very complicated both in time and space, so there is a need for a new approach.

HTPM algorithm improved the Apriori-like process first by trying to find all the length-1 frequency events in a first database scan, and using these proven frequent events to generate candidates. This prevents candidates to be selected among non-frequent events. Secondly, HTPM generates new candidates by two patterns consisting of common events only. This step controls both the number and length of the new pattern generated. The detail steps are represented as follows:
1) Find all the frequent events by scanning the database (Table 1) once and put these patterns into a list. This step actually converts the database from a transaction-oriented to event-oriented database, and purges the non-frequent events at the same time. All the events are kept in the example from thsi paper because of the number of occurrences are above the minimum support threshold.
2) Combine each two patterns form the list in step (1), and get a new list of patterns (Table 2):
In step 1, three patterns are kept: $a^+ <a^-$, $b^+ <b^-$ and c. When combining each two patterns, their occurrence record is joined, and each transaction will generate a new pattern. Count the patterns of all

Table 1: Example Database

ID	Event	Time
1	C	6
1	C	8
1	A	[5,10]
1	B	[6,12]
1	A	[8,12]
2	C	6
2	C	8
2	B	[6,11]
2	A	[8,11]
3	C	4
3	A	[4,10]
3	B	[4,12]
3	A	[9,12]

Table 2: Event oriented database

L1	Index	Occurrence
$a^+ < a^-$	S1	(5,10),(8,12)
	S2	(5,10),(8,12)
	S3	(5,10),(8,12)
$b^+ < b^-$	S1	(5,10),(8,12)
	S2	(5,10),(8,12)
	S3	(5,10),(8,12)
c	S1	(6),(8)
	S2	(6),(8)
	S3	(4)

transactions, and keep the one larger than the minimum support threshold only. Take pattern $a^+ < a^-$ and c for example, we get three candidate patterns by joining pattern $a^+ < a^-$ and c (Table 3): only patterns $a^+ < c < a^-$ and $c < a^+ < a^-$ are saved.

Table 3: Possible candidates

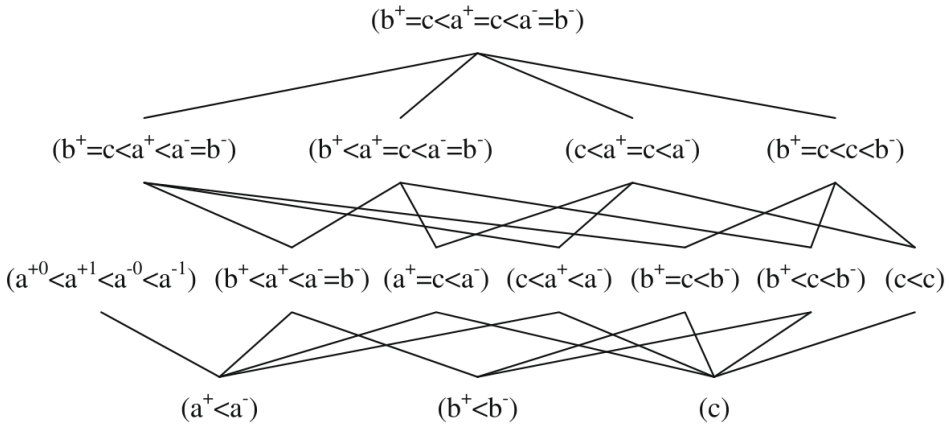
Candidate	$\mathbf{a^+ < c < a^-}$	$\mathbf{c < a^+ < a^-}$	$\mathbf{a^+ = c < a^-}$
1	(5,6,10),(5,8,10)	(6,8,12)	(8,8,12)
2		(6,8,11)	(8,8,11)
3		(4,9,12)	(4,4,10)

3) Redo step (2) and only combine two patterns until no more patterns are found. This process will keep going and get a result set shown in figure 1:

3.2 System Design

The system structure for implement the algorithm is not difficult to get, as follow the sequence of algorithm design. Few issues still need to be addressed for simplicity and complete system design, one is the storage of database, another is the structure of pattern.

Figure 1: The result set of HTPM on example dataset



The event object is implemented by an new class rather than object for flexibility and future use. The input database can be represent by a level−2 list List<List<Event>> db = new ArrayList<List<Event>>(). As we know, pattern consists of events and is formed by a certain relationship. So the first option is using event object represent pattern. But there is a problem which is how to represent the sequence of events. Obviously, events have to be broken for represent pattern, this makes the program more complicated than before. The second option is to divide each event like $a^+ <a^-$ to individual elements like a^+ and a^-, and using these elements to form a pattern. This is more flexible but extra mechanism has to be involved to make sure each element shows as a pair. So finally, a third option comes with despite of the event object, the sequence of pattern can be reformed when it was created and represented by string. Similarly, all the occurrence records need to store in the pattern object. This option may cost a lot memory, but this is still the most understandable option comparing to others.

3.3 Concurrent Programming

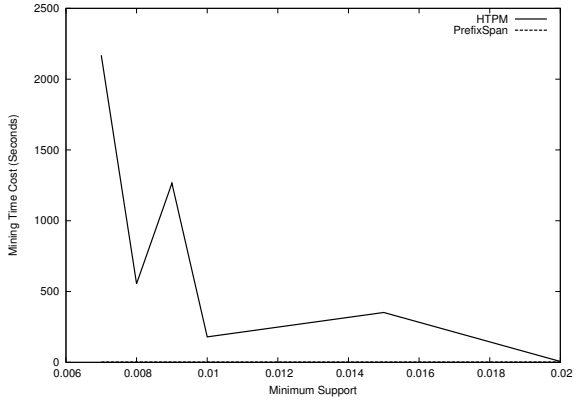
Concurrent programming means using more than one CPU when available. With the development of CPU, concurrent programming became very important to make full use of the system resources.
In JAVA programming language used for the implementation of the system described in this paper, multi-thread technology is employed to spread tasks to different CPU cores. Each thread has a life cycle control by the JAVA Virtual Machine (JVM). The main thread created by the main function is going to be the first and last thread alive when the program is running.
As the basic concept of concurrent programming follows the strategy of divide and conquer, the HTPM can only following this by calling the Join() function concurrently. Each Join() function is called with different parameters, so there is not conflict with other threads.

3.4 Different Candidate Generate Policy

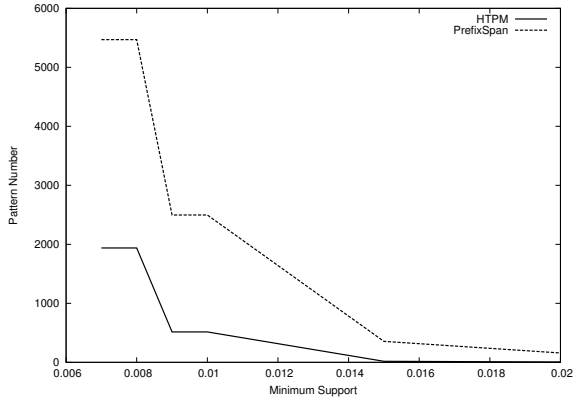
HTPM employs a process of using existing patterns to form new patterns by merging the occurrence time of the existing patterns. Obviously, with increasing the length of patterns, the occurrence record becomes smaller, and reduces the possibility to form a new pattern. The only way to increase this possibility of generating new patterns is finding more occurrence records.
The first way to keep as many occurrence records as possible is to reduce the steps of forming new patterns. Take pattern $b^+ < c < a^+ = c < a^- = b^-$ for example. In the algorithm before, this pattern was formed by pattern $b^+ = c < a^+ < a^- = b^-$ and pattern $b^+ < a^+ = c < a^- = b^-$, and we used four steps to find this pattern. But if we join pattern $b^+ < a^+ < a^- = b^-$ and pattern c<c in level two, we will also get pattern $b^+ < c < a^+ = c < a^- = b^-$ in the third step.
This means each two of the patterns in the same level should be tried to be joined together, and see if

Figure 2: HTPM vs Prefixspan Using the IBM Data

(a) Time consumption of HTPM and PrefixSpan based on IBM Data



(b) Number of Patterns Found by HTPM and PrefixSpan based on the IBM Data



there is a new pattern formed. However, this process has a shortcoming to require an extra process to inspect if the new pattern created is duplicated or not.
While keeping on finding more available occurrence records in HTPM, we found the only level containing more occurrence records is level-1, which are patterns consisting of a specific single event. So we come up with a second way to form new patterns by keeping using the level-1 patterns. Theoretically, this method will continue to generate patterns until possible.

4 Simulation Results

After testing the program with simple example data, a more complicated simulation data will be used.
The first simulation data is generated by the IBM data generator, which generates configurable datasets emulating user transactions. The data generator can manually set up the data attributes such as pattern number, pattern length and pattern density.
The second set of data is simulation data but includes interval-based events. This generator can let user define patterns and occurrence possibilities. Event name is followed by start time and end time. Each transaction ends with event name -101.
Next this section evaluates the performance of HTPM and analyses how HTPM can be improved by adding different combine policies and programming techniques. The following tests are performed, and in each test, both running time and pattern number found are compared. All the experiments are running on a machine with an Intel core 2 CPU with 2G RAM and using Java SE 7u25.

4.1 HTPM vs PrefixSpan using the IBM data

PrefixSpan is a well-known data mining algorithm with good performance, so it is meaningful to compare the result of HTPM and PrefixSpan.
For the IBM data, six different minimum support thresholds are tested from 0.02 to 0.007. Both running time and the discovered frequent pattern numbers are compared, as shown in figure 2.

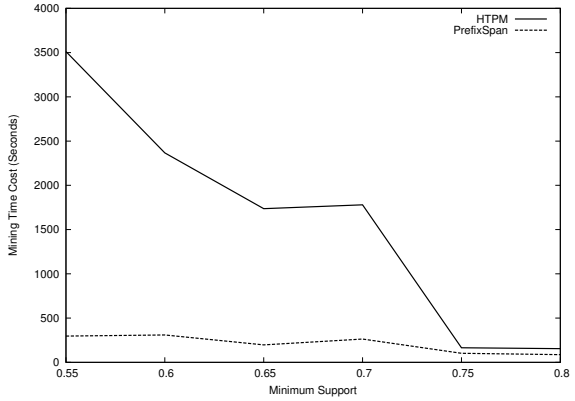
For the data generator dataset, because of high data repeat rate, the support level is rising up to $0.8-0.55$. Also, both running time and the number of frequent patterns found are compared, as shown in figure 3.

4.2 HTPM with different combine policies when using the IBM data

As two different combination policies are proposed before, each of them is tested with two different datasets. For the data resulted from the data generator, it also can be tested by mining interval-based

Figure 3: HTPM vs PrefixSpan Using the Synthetic Data

(a) Execution Time of HTPM and PrefixSpan based on the Synthetic Data



(b) Number of Patterns Found from HTPM and PrefixSpan based on the Synthetic Data

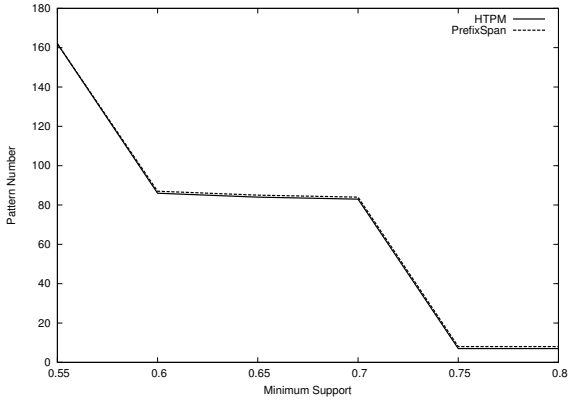
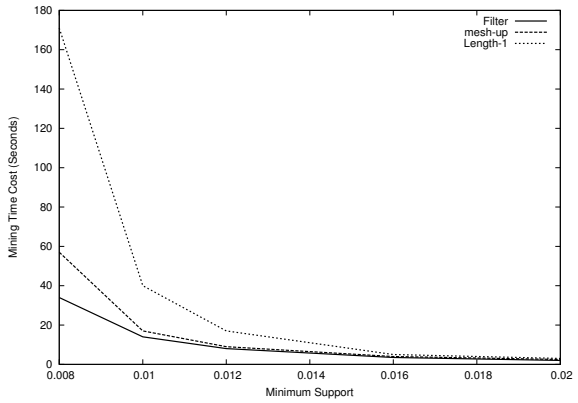
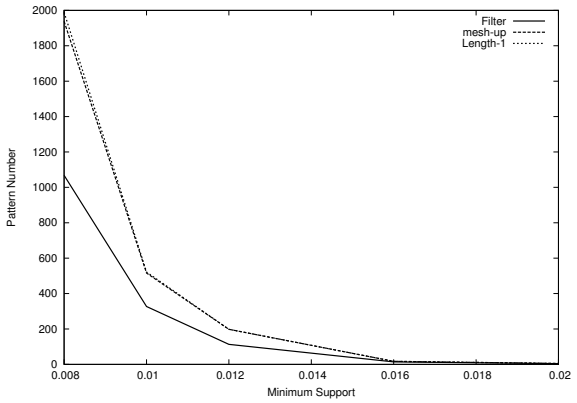


Figure 4: Three Different Combine Policies

(a) Execution Time for Different Combine Policies with the IBM Data



(b) Number of Patterns Found for Different Combine Policies with the IBM Data



patterns.

As shown in figure 4, ”filter” means using the combination policy proposed by the HTPM author, join two patterns only if they share same events in common; ”full-mesh” means join each two patterns in the same list; and ”length-1” means always form new pattern by joining patterns in the first level, which means have the most occurrence records.
For the IBM data, it can only be treated as point-based events. The minimum support level ranges from 0.02 to 0.008. Both running time and pattern numbers are compared.

4.3 HTPM with different combine policies using interval-based data generator data

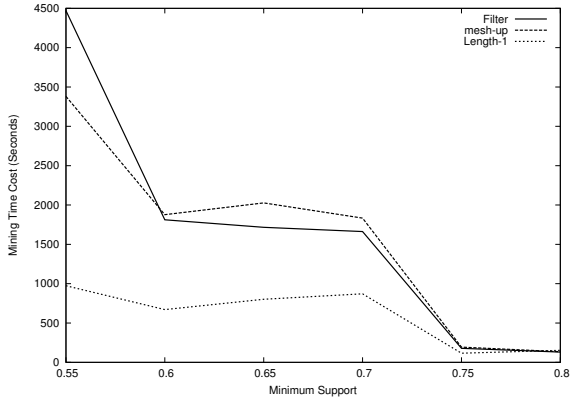
Data generator dataset can be parsed as point-based events or interval-based events. Both of them are tested, with minimum support level ranges from 0.8 to 0.55. Both running time and pattern numbers are compared, as shown in figure 5.

4.4 Single- and Multi-Thread HTPM Performance Comparison

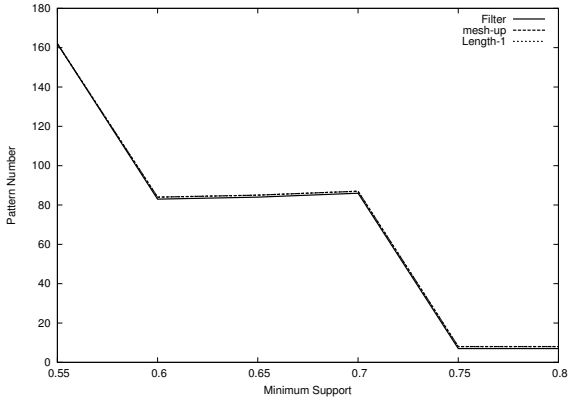
The performance of classic HTPM implementation and HTPM implementation using concurrent programming methods are compared in this subsection. Multi-thread is a general way to improve program

Figure 5: Three Different Combination Policies

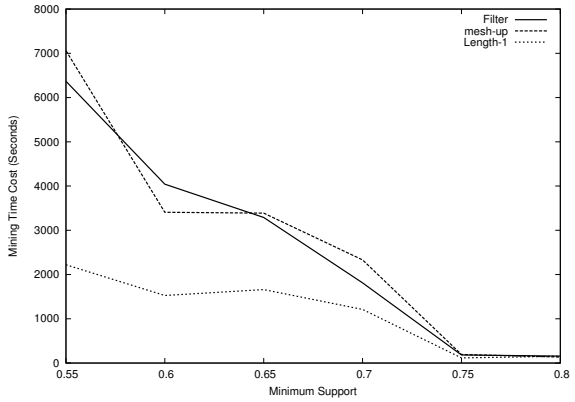
(a) Execution Time for Different Combination Policies with the Point-based Synthetic Data



(b) Number of Patterns Found for Different Combination Policies with the Point-based Synthetic Data



(c) Execution Time for Different Combination Policies with the Interval-based Synthetic Data



(d) Number of Patterns Found for Different Combination Policies with the Interval-based Synthetic Data

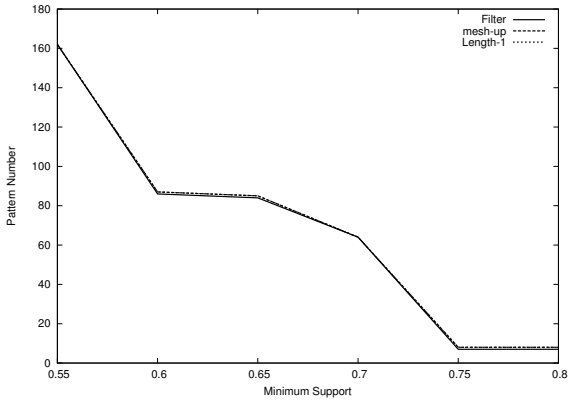
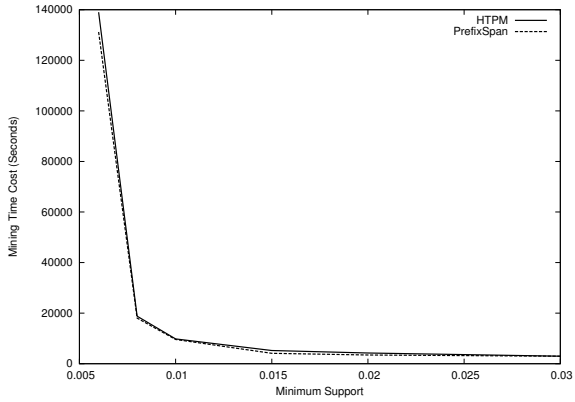
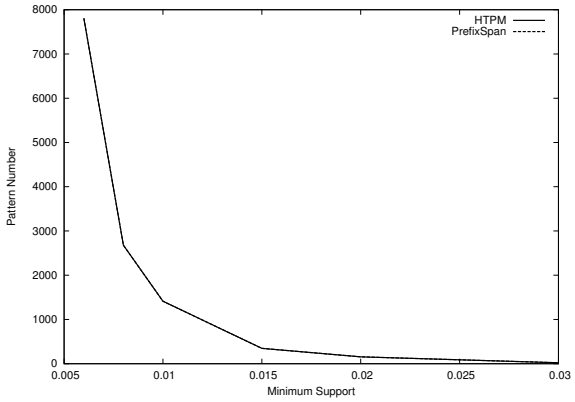


Figure 6: Multi-Thread HTPM

(a) Execution Time for Single- and Multi-thread HTPM



(b) Number of Patterns Found for Single- and Multi-thread HTPM



performance by allocating tasks to different CPUs.

In this test, the IBM data only is used for comparison because of low frequency data is more easily controlled by program and there is a lower error rates. Both running time and pattern numbers are compared, using the minimum support level from 0.03 to 0.006, as shown in figure 6.

5 ANALYSIS

This section analyses the data produced in last section.

5.1 HTPM vs PrefixSpan

HTPM and PrefixSpan find approximately the same number of patterns, but HTPM takes more time than PrefixSpan, both in the IBM dataset and the Data generator dataset.
This can be explained as although HTPM can be used for mining point-based patterns, it still has to take extra time to inspect if each event has an end time. However, the PrefixSpan approach does not care about the time label of each event at all.
Secondly, different from the top-down approach of PrefixSpan, HTPM uses existing patterns to generate candidates, which means the load is exponentially increasing while more patterns are coming. However, the processing task in PrefixSpan keeps decreasing as a projection method is used.
Furthermore, the PrefixSpan is not affected by duplicate patterns because only one event is found in each round; but the HTPM needs an extra effort to keep no duplicate events (same name and same occurrence time) occurring in each pattern.
All these factors explain the larger processing time of the HTPM algorithm. So even if HTPM can be used for processing pure point-based event, it is still a better option to use PrefiSpan or other point-based event oriented algorithms in terms of performance.

5.2 HTPM Combine Policy

Two different datasets and both point-based and interval-based events are tested on three different combination policies. The experiment based on the IBM dataset shows the filter policy which was proposed by HTPM author resulted in the least time but also found the least patterns. The mesh-up and length-1 combine policies discovered the same number of patterns, but the length-1 method required longer time, especially with lower minimum support level.
It is not surprised that the mesh-up and length-1 methods have found more patterns. As we discussed before, both these two methods use patterns which contain more occurrence records, which means more new patterns could be discovered in terms of more candidates to be formed. More patterns means higher processing time, which is corresponds to the result discussed before.
The result got a little different when processing the data generator dataset. Both methods found exactly the same number of patterns, but length-1 method required less time. This can be explained by the attributes of the datasets. The datasets generated have three whole patterns, and each pattern can be divided to shorter length smaller patterns. Each whole pattern is fully contained in each transaction, which means the occurrence record did not reduce while the pattern length grew. But length-1 pattern growth policy forms new pattern by adding only one event at each round and that means less load to process when generating new candidates. Thia explains the result obtained.
In conclusion it can be said that the original filter method takes lower time to find certain number of patterns, and the new length-1 solution takes the most time to find more patterns. This makes a trade-off between time and pattern numbers, which can give user more flexible options while processing data mining.

5.3 Multi-Thread on HTPM

The result shows that both single thread HTPM and multiple threads HTPM find exactly the same number of patterns, and the multi-thread HTPM did not reduce the processing time very much.
This does make sense after reviewing the process of HTPM and the concept of multi-thread. The basic concept of concurrency programming is to divide program into small tasks and alocate them to different CPUs. Different from PrefixSpan, the HTPM process is really straight forward: each set of patterns found will be used to make a new set of patterns. So the only process which can be divided is the part of

finding each candidate. However, with hundreds existing patterns, the candidate form threads can be in the order of thousands. Extra effort is required to control these threads, which means the program does not get very much benefit from a multi-threaded approach.

6 Conclusion

In this paper, a hybrid-based pattern mining algorithm HTPM has been evaluated when dealing with different datasets and event types. After implementing each function of the algorithm, some new pattern combine approaches are proposed and tested to see if the performance of the algorithm can be further improved. One of the improvement aims to reduce the mining time by using multi-threading which makes full use of the existing multi-core environments. Although the benefit when using with HTPM was not evident, it can still provide an idea to improve mining performance for other algorithms. The second improvement is using different candidate generation policies to find more patterns. With more patterns found in the datasets, the users could get a better understanding of the dataset, therefore make a better decision.
This work can be extended in several ways. First, the complexity of the algorithm is still very high comparing to PrefixSpan and other point-based pattern mining algorithms. More efforts could be done to improve the algorithm, both in terms of performance and efficiency. Secondly, this algorithm can be applied in different scenarios, including stock market analysis. The potential applications in various areas could help users benefit from the benefits of data mining.

7 Acknowledgements

This work is funded by Enterprise Ireland Innovation Partnership Programme with Ericsson Ireland under grant agreement IP/2011/0135 [Dublin City University and Ericsson,].

References

[Chu and Lin, 2005] Chu, W. and Lin, T., editors (2005). *Foundations and Advances in Data Mining*, volume 180 of *Studies in Fuzziness and Soft Computing*. Springer-Verlag, Berlin/Heidelberg.

[Dublin City University and Ericsson,] Dublin City University and Ericsson. E-Stream Project.

[Fan et al., 2011] Fan, S. X., Yeh, J.-S., and Lin, Y.-L. (2011). Hybrid Temporal Pattern Mining with Time Grain on Stock Index. In *2011 Fifth International Conference on Genetic and Evolutionary Computing*, pages 212–215. IEEE.

[Han et al., 2007] Han, J., Cheng, H., Xin, D., and Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86.

[Han and Kamber, 2006] Han, J. and Kamber, M. (2006). *Data Mining, Southeast Asia Edition: Concepts and Techniques (Google eBook)*. Morgan Kaufmann.

[Hirate and Yamana, 2006] Hirate, Y. and Yamana, H. (2006). Generalized sequential pattern mining with item intervals. *Journal of computers*, 1(3):51–60.

[Hu et al., 2009] Hu, Y.-H., Huang, T. C.-K., Yang, H.-R., and Chen, Y.-L. (2009). On mining multi-time-interval sequential patterns. *Data & Knowledge Engineering*, 68(10):1112–1127.

[Kam and Fu, 2000] Kam, P.-s. and Fu, A. W.-C. (2000). *Discovering temporal patterns for interval-based events*. Springer.

[Kambayashi et al., 2000] Kambayashi, Y., Mohania, M., and Tjoa, A. M., editors (2000). *Data Warehousing and Knowledge Discovery*, volume 1874 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Lakshmi and Raghunandhan, 2011] Lakshmi, B. and Raghunandhan, G. (2011). A conceptual overview of data mining. In *2011 National Conference on Innovations in Emerging Technology*, pages 27–32. IEEE.

[Mannila, 1996] Mannila, H. (1996). Data mining: machine learning, statistics, and databases. In *Proceedings of 8th International Conference on Scientific and Statistical Data Base Management*, pages 2–9. IEEE Comput. Soc. Press.

[Muntean C. H., 2001] Muntean C. H., McManis J., M. J. (2001). The influence of web page images on the performance of web servers. *IEEE International Conference on Networking*, pages 821–828.

[Nisbet et al., 2009] Nisbet, R., IV, J. E., and Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications (Google eBook)*. Academic Press.

[Pei et al., 2001] Pei, J., Pinto, H., and Chen, Q. (2001). Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. *2013 IEEE 29th ...*.

[Villafane et al., 1999] Villafane, R., Hua, K. A., Tran, D., and Maulik, B. (1999). Mining interval time series. In *DataWarehousing and Knowledge Discovery*, pages 318–330. Springer.

[Wu and Chen, 2007] Wu, S.-Y. and Chen, Y.-L. (2007). Mining Nonambiguous Temporal Patterns for Interval-Based Events. *IEEE Transactions on Knowledge and Data Engineering*, 19(6):742–758.

[Wu and Chen, 2009] Wu, S.-Y. and Chen, Y.-L. (2009). Discovering hybrid temporal patterns from sequences consisting of point- and interval-based events. *Data & Knowledge Engineering*, 68(11):1309–1330.

[Xiaodan, 2011] Xiaodan, Z. (2011). Plain discussion of data mining technology research. In *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 296–298. IEEE.

[Zaki, 2001] Zaki, M. J. (2001). Spade: An efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2):31–60.

Telecom Network Performance Analysis Using Big Data Technologies

Zhuo Wu, Faisal Zaman and Gabriel-Miro Muntean

Dublin City University, Glasnevin, Dublin 9
zhuo.wu3@mail.dcu.ie, (faisal.zaman, gabriel.muntean)@dcu.ie

Abstract

The advent of smart telecommunication technologies coupled with constant increase of the users has increased the number of network elements to be managed and monitored in exponential scale. Combination of these, result in an explosion of network management data and adding further complexity for the network operators. In this paper, we have designed a two-staged analytical framework for analyzing the telecom network performance. The network traces are processed in a distributed way in the first stage and in the second stage the event traces are reduced based the complex relationship between network entities over different incidents. In the framework, a set of data approximation methods are integrated to run on the big data technique Hadoop and with the parallel computational ability of MapReduce the data approximation methods are able to analyze considerable volume of network traces.

Keywords: Telecom network, Data Analysis, Hadoop, MapReduce

1 Introduction

The telecommunications industry is generating massive amounts of data. A Cisco report [Cisco, 2013] shows how world-wide mobile data traffic is increasing at an astounding rate and this growing trend is set to continue in the foreseable future. For telecom operators, this large amounts of incoming data (i.e. big data streams) have both positive and negative conseqiuences. Naturally, there is much useful information in these streams in form of certain patterns, which can be harvested and used for the improvement of the overall system's performance. One of the most important tasks in network analysis is to discover these patterns from the historic data, identify them in the current data stream and also predict their occurrence in future data. Data mining tools can be employed for data processing and the large amounts of data provide a representative sample, but at same time increase the difficulty of the network analysis.

These difficulties faced by researchers when employing traditional methods of data mining for big data processing are caused by the fact that the amount of data is so large, that often exceeds the system's memory storage capacity. Gigabytes or even terabytes of data is considered normal, but it is hard to find the required large memory space on normal machines. Additionally, running data analytics algorithms might also fail due to out of memory errors. The second problem is long execution time when datasets become extremely large, as mentioned in [Chen, 2010, Wei Gao et al., 2011, Butalia et al., 2008]. The value of data is time-sensitive; for example, if the daily amount of data produced is 1TB, but only half of that can be processed in the same period of time, the telecom operator cannot get the results it expects every day. The out-of-date results are less meaningful for operators and therefore they require a solution that can efficiently manage and process the rising volume of data.

Simple light weight data projection techniques [Bingham and Mannila, 2001] can serve the purpose of reducing the data dimension, but is capable of handling limited volumes of data only. Parallelizing the projection can solve the problem, but there is a possibility of higher approximation error. Further reduction can achieved through extracting the events linked through some relation e.g., temporal or spatial

co-occurrence and contextual relation. From a data reduction point of view, events can be discarded after the failure to exhibit the pre-defined relational structure. Utilizing the heuristic of correlative structure between the events can efficiently reduce events [Zaman et al., 2014].

The data projection parallelization can be designed to work in conjunction with Hadoop-MapReduce [Dean and Ghemawat, 2008], which is an ubiquitous parallel computing paradigm to process large data. MapReduce integrated with a distributed computing technique - Hadoop, which is able to process large datasets and maintain a higher approximation error. Hadoop was originally developed and used by large Internet companies like Yahoo and Facebook and is being widely used in other areas such as telecom, finance, government and so on [Cloudera, 2010].

This paper describes the Hadoop-MapReduce paradigm for processing extensively large telecom network traces in order to gain a deep insight into network performance and eventualy be able to take performance-improving actions. Some data mining algorithms are re-designed to fit in the MapReduce parallel programming model context and are are run on a Hadoop cluster to utilize the advantages of distributed computing. The contributions of this paper are as follows:

- Evaluate the efficiency of processing network trace files in a Hadoop cluster.
- Implement very sparse random projection to reduce the trace files using MapReduce.
- Implement the correlation matrix calculation to discover correlated events.

This paper is structured as follows: Section 2 introduces the technical approach used to process the data and the network trace files used in experiments. Section 3 describes the implementation of the approach using Hadoop-MapReduce. The results of experiments are presented in Section 4 and the analysis of the results is made in Section 5. In particular it is tested how the number of nodes affects the algorithms' performance on Hadoop, and how different parameters influence the correlation between events. The paper is concluded in Section 6.

2 Approach

This section describes the technical approach used to process the data as part of the proposed analytical framework and the network trace files used in our experiments. Figure 1 illustrates the overall framework of our work and its major components.

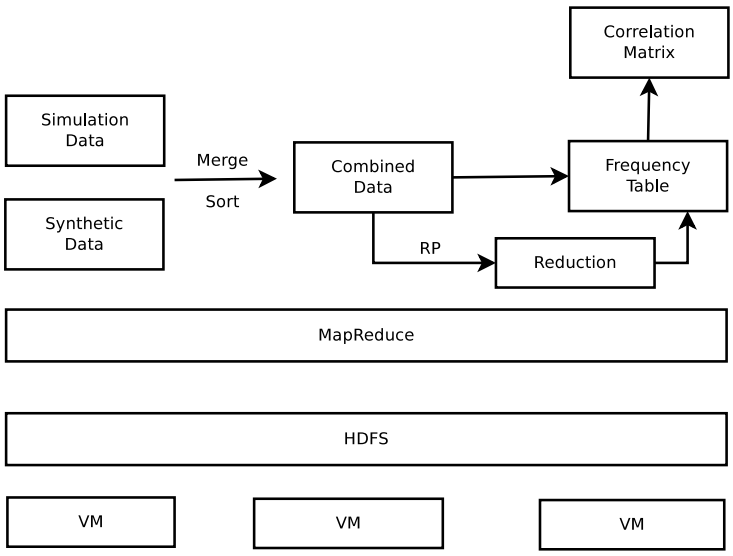


Figure 1: Overall Hadoop-based big data processing framework

In the Hadoop cluster considered, there are three virtual machines (VM) as the distributed computing environment. The trace files are stored in the Hadoop Distributed File System (HDFS) and the code runs

on the MapReduce component. In our framework, the trace files are merged and sorted, then a frequency table is generted and finally the correlation values are computed. We apply the sparse random projection (RP) on the trace files to reduce the amount of data and check the effect on the result analysis.

2.1 Dataset Description

The network trace files used in this work are synthetically generated, are acquired from two different sources and are combined. One part of the input data comes from the logs of a network simulator. Events traces were generated using LTE-Sim [Piro et al., 2011], which is as an event-driven simulator written in C++. LTE-Sim is open-source and includes both the Evolved Universal Terrestrial Radio Access (E-UTRAN) and the Evolved Packet System (EPS). In particular, LTE-Sim provides key features such as flow management (i.e. radio bearers, QoS requirements, etc.), feedback control, multiple types of downlink schedulers, etc. LTE-Sim was installed in an Ubuntu 12.10 linux system. The advantage of simulation data is that it is very close to the data from real life situations. A similar but more advanced technique is deployed in [Yuan et al., 2014] to mimic the network behaviour. However for our big data analysis, it takes too much time to generate large amounts of information. An additional problem is that we do not know the patterns in the simulation logs, which affects the result validation. We have added a second source of synthetic data in order to alleviate this problem.

The second part of input data is artificially generated by an event pattern generator. It is able to generate large traces with low latency. In the generator, the events are specified with an end-time, in addition to a start time. The event duration (end time$\sim$start time) of an event instance is drawn from a uniform distribution, with a specified minimum and maximum value for each event type, specified separately for each pattern type and separately for each event type in the background noise. The event pattern definition specifies the statistical relationship in time between the different events that comprise a pattern. It is specified as two matrices, one storing a set of distribution mean values and one storing variance values. The $[i, j]$th entry in the mean matrix gives the mean time from the occurrence of a type i event in a pattern instance until the occurrence of a type j event in the pattern instance.

The raw trace data is in format of event logs, and are stored in two files. Each line in the file is a record including when and what event happens. To apply data mining tools to the combined data, the data needs to be normalized. Merging the two data files and then sorting the events according to their timestamps achieves the desired normalization.

2.2 Hadoop Distributed File System

In order to utilize Hadoop, the raw data is stored on HDFS. The HDFS, derived from the Google File System [Ghemawat et al., 2003], is the core file system for storing large datasets in a distributed computing environment. There are usually several nodes in one Hadoop cluster. The data on HDFS is split into small blocks and the blocks are stored in these nodes. For reliability, the data is duplicated and several copies are stored on HDFS.

The HDFS uses a master/slave architecture. It has two types of nodes: `NameNode` and `DataNode`. The NameNode is the master and the DataNode is the slave. The NameNode keeps track of how the files are split and stored in the file system. All the data blocks are stored and processed in the DataNode.

The HDFS provides fault tolerance when running tasks. If a DataNode fails, the NameNode can detect this and re-schedules all the tasks on the failed node. If a NameNode fails, the cluster will be down and all the DataNodes will stop working. Users have to restart it and restore the operation from the backup.

2.3 MapReduce

The MapReduce is a programming model [Dean and Ghemawat, 2008] which describes solutions by means of two functions:

1. `Map()`

2. `Reduce()`

Splitting the parallel processing model in two phases increased the efficiency. In `Map` phase the data are allocated with key values in order to distribute to several processors and in the `Reduce` phase data with common key values are parsed. In this way MapReduce can process considerable amount of data with low latency.

Figure 2 shows the process flow of MapReduce in the experiments. At the beginning, the raw data is split into small blocks to be processed; for instance four blocks are illustrated in the figure, following the split. The mapping phase plays a role in filtering and sorting the data. Each record of data is mapped into a collection according to a key, which is called `shuffle`. In the shuffle and sort phase, all mapping outputs are sorted in collections based on the keys. In the Reduce phase, all records in each collection are reduced and a summarized result is produced.

The MapReduce provides an easy way to produce parallel distributed programs. In order to run on Hadoop, the code should be modified to correspond with the MapReduce model. Hadoop is responsible for executing MapReduce programs, assigning tasks to slave nodes and running in a parallel distributed way. Programmers are able to concentrate themselves on coding in MapReduce manner, caring little about running programs in parallel.

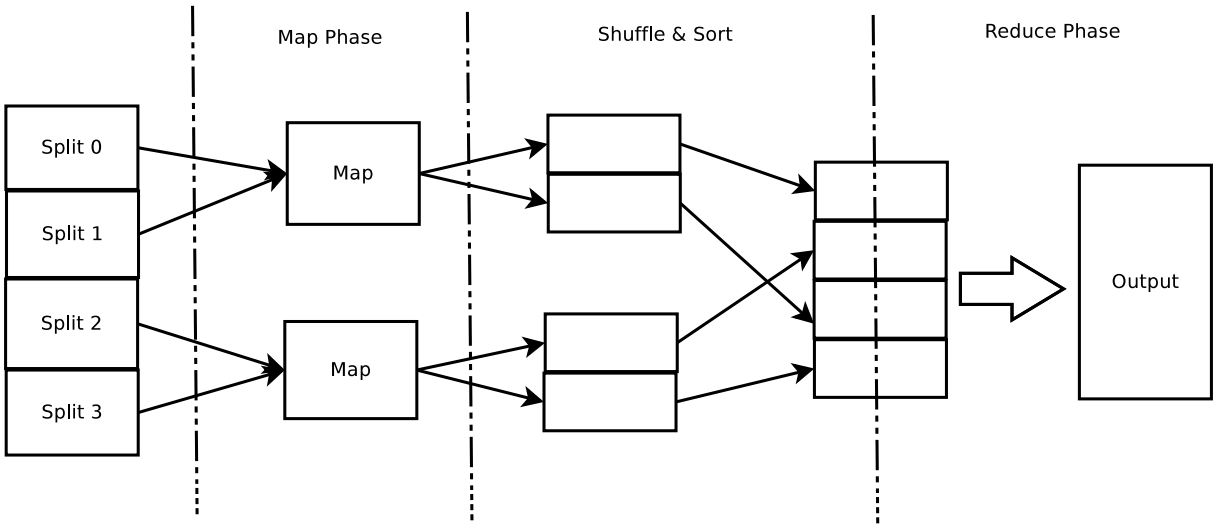


Figure 2: The MapReduce programming model

2.4 Virtualization

The virtualization technology can simulate real computer environments. For example, the size of memory, the disk storage and network connection can be configured via software. With virtualization, we can install Hadoop runtime environment on virtual machines instead of physical ones [Xu et al., 2012]. The advantage of virtual machines is that we can deploy several nodes in one computer, simplifying the deployment.

Among the many virtualization products available, we chose Oracle VM VirtualBox for our experiments. It is a powerful x86 and AMD64/Intel64 architecture which is able to support the Ubuntu operating system. The key point is that it is totally free for both individual users and enterprises.

3 Implementation with MapReduce

This section describes the implementation of our approach using Hadoop. In this section we describe the functionalities of the framework in terms of the `Map` and `Reduce`. The functions are modified in order to follow the MapReduce model.

3.1 Merge and Sort

Both merging and sorting data can be executed in one MapReduce program and these are one of the basic tasks operated using MapReduce. As the formats of simulation data and synthetic data are different, we need to pre-process the two sets of input data before merging and sorting. The pre-processing operation includes the following two steps.

1) For the simulation data set, mapping the Event Identifiers (EventIDs) to the 0 to 11 range, is done according to equation (1):

$$f(x) = \begin{cases} \frac{x+1}{2} \mod 11 & , x \leq 2200 \\ 11 & , x > 2200 \end{cases} \tag{1}$$

Note that, EventIDs greater than 2200 represent cell congestion events and are mapped as EventID 11. Using this equation, other events with ID lesser than 2200 are mapped to EventIDs 0 to 10. The detection of correlative structure between events will leverage on the mapped events.

2) Forthe synthetic data set, shifting the EventID values to make them distinct from those in the simulation data set is performed by using equation (2):

$$f(x) = x + 20 \tag{2}$$

The implementation of merging and sorting data is straightforward. In the Map phase, the timestamp is employed as the key of each record so that all records can be ordered based on the timestamp. In the Reduce phase, the records are printed.

3.2 Frequency Table

The event logs, which record event occurrences over time, can be regarded as time series data. They have a natural temporal ordering. For analysis, we partition the data using a small timeslot and count event frequency in each timeslot. By listing all timeslot information, a frequency table is produced.

The frequency table can be regarded as a large matrix that records what events happen in each time slot. In the matrix, each row represents a timeslot and each cell marks how many times the specified event happens.

For the large matrix, we employ two formats to represent it. The first one is using the CSV format, in which each line represents a row of matrix, and the cells in each row are separated by commas. The CSV format records all values in the table and is convenient for human reading. In reality, the matrix is a sparse one which means most items are zero. The major disadvantage of the CSV format is that it takes much space to store the many zeros. The other approach employed is a key-value Key-Value (KV) format, in which each line represents an element of the matrix. The key contains the row and column numbers, separated by comma. The value is the real value of the element in matrix. The separator between the key and value is a tab.

3.3 Random Projection

The random projection [Bingham and Mannila, 2001] is an approach to reduce the dimension of the data. We utilize random projection to find a smaller matrix, which contains few number of columns than the original matrix, but maintains all interesting components of the original matrix, such as all event patterns. The analysis on the resulted narrow matrix is much more efficient. Additionally, if the narrow matrix is small enough to be easily kept in the normal memory, traditional data mining methods can be applied for processing.

To apply random projection, the original matrix should be multiplied by a random matrix. Suppose that the original matrix A is a matrix of size N rows by M columns ($N \times M$). The processing steps include:

1. Generate a random matrix P of size $M \times K$

2. Multiply random matrix with original matrix as: $B = AP$

Several approaches were investigated to generate the random matrix and the Ping Li's very sparse random projection matrics was chosen [Li et al., 2006]. The random matrix is generated by the very sparse distribution from equation (3):

$$R_{ij} = \sqrt{s} \cdot \begin{cases} 1 & , prob. \frac{1}{2s} \\ 0 & , prob. 1 - \frac{1}{s} \\ -1 & , prob. \frac{1}{2s} \end{cases} \tag{3}$$

, here s controls the sparsity of the matrix and $prob$ represents the probability of the random values.

The advantage of this approach over sparse projection matrix [Achlioptas, 2003], is that with high sparsity factor (s) the approximation accuracy of the projection is considerably high. In addition to this, the very sparse matrix reduces the computational complexity of the approximation process significantly.

To generate a very sparse random matrix, we need to specify the size and the sparsity of the matrix. The size is the number of rows and columns in the matrix. The sparsity, which is the value of s in the very sparse distribution, means that only $\frac{1}{s}$ of elements in the matrix are non-zeros.

For a very sparse distribution, there is no theoretical guideline over choosing an optimal value of sparsity but in the paper [Li et al., 2006] the authors suggest that a small accuracy reduction is introduced when the sparsity s is larger than 3.

To implement the very sparse random projections, we used the KV format to store the two matrices. The matrix multiplication $P = MN$ is defined as in equation (4):

$$B_{ik} = \sum_{j} A_{ij} P_{jk} \tag{4}$$

The implementation on Hadoop can be divided into two stages and needs two MapReduce models to be described. In the first stage the following functions are employed:

- `Map()`: For each matrix element A_{ij}, generate the $\texttt{key}_{\langle i,n,j \rangle}$, where $n = 0 \ldots k$; For each matrix element P_{jk}, generate the $\texttt{key}_{\langle n,k,j \rangle}$, where $n = 0 \ldots i$
- `Reduce()`: Combine the values with same key if the number of values is 2.

In the second stage:

- `Map()`: Write the input to the output.
- `Reduce()`: Sum up the values with the same key.

In our work, we used random projection to reduce the amount of data. Suppose the number of event logs is M and they contain N event types. The algorithm is described as follows:

1. Create an $N \times M$ matrix to store the information of event logs. The columns of the matrix represent the EventIDs and the rows represent the timeslot (each timeslot is a pre-defined non-overlapping temporal interval). Initially, all elements in the matrix are 0s.
2. Transform the event logs into a frequency table. Read the event logs line by line and set the value at specific EventID cell to 1 according to the following: matrix$[i][j]$ = 1, where i is line number, j is EventID
3. Now for each EventID create an $M \times 1$ matrix using very sparse random matrix. For all the EventIDs in this way a projected matrix of size $M \times K$ is generated.
4. According to the sparse matrix resulted in Step 3, some lines can be deleted from the event logs. The operation is described as follows: `if matrix`$[i][0] == 0$ `then delete logs`$[i]$

3.4 Correlation

The correlation indicates the relationship between two measured variables. High value of the correlation coefficient means there is strong relationship between the two measured variables. The correlation can be calculated using the Person's correlation coefficient as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{5}$$

Correlation is applied on the sparseley projected matrix in order to find the correlated EventIDs. The computation of a correlation matrix is implemented in MapReduce framework realizing the following steps:.

Step 1: Standardize matrix A by column. This step consist of 3 jobs: **division**, **deviation**, **averaging**. Each matrix element is divided by the standard deviation after the deviation from average.

Suppose $A = [A_1, A_2, \ldots A_n]$, A_i is the column vector of matrix A. The standardized matrix $C = [C_1, C_2, \ldots C_n]$ and we can compute the C_n by:

$$C_n = \frac{A_n - \overline{A_n}}{\sqrt{\frac{1}{N}(A_n - \overline{A_n})^2}} \tag{6}$$

Step 2: Compute the correlation matrix using a **multiplication** job:

$$R = \frac{1}{N}C^T C \tag{7}$$

The following figure shows the flow chart of these jobs.

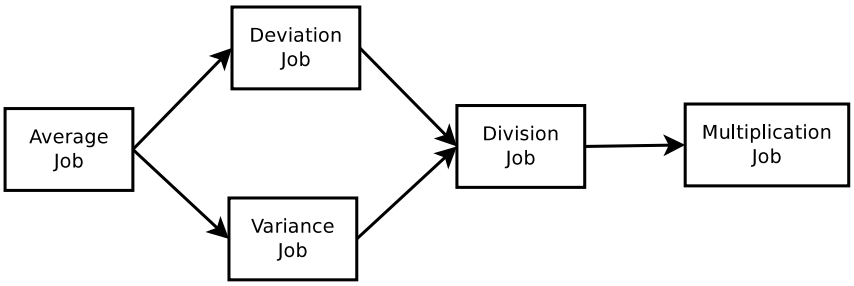


Figure 3: The flow chart of correlation jobs

4 Experiments and Results

This section describes several experiments we designed for testing our analysis approach and evaluating the performance using Hadoop. For our experiments, we used virtual machines instead of physical ones. Since Hadoop is able to run on regular machines [Patel et al., 2012], we deliberately chose a lower configuration for the experiments. For each node, we only allocated 1024 MB of memory and 20.0 GB of disk storage. The network adapter is a bridged adapter, so that each node has a static IP address.

In the simulation data, we defined 50 cells to accept User Equipment (UE) and each cell can hold maximum 10 UEs. There are continuous UEs joining and leaving cells. Both the time and cells are random. The EventIDs represents the number of UEs in a cell and if it reaches the maximum number, the cell will lead to a *cell-full error* and empty all the UEs in cell.

In synthetic data, we chose 80 types of events whose IDs range from 0 to 79. When generating the synthetic data, we defined a pattern that EventIDs are from 0 to 9. After combining data, the EventIDs

of synthetic part are shifted to 20∼79. The first experiment tests merging and sorting of data. At same time, analysed the performance using different numbers of nodes.

There are two files to be merged and sorted. One includes the simulator data with 28,749 events and the simulator took about 10 hours to generate this. The other one contans the synthetic data and includes 263,784 events and it only took seconds to generate it. The time duration of events is the same and the two files size are 0.8MB and 1.3 MB, respectively. We executed the program using 1, 2, 3 nodes and recorded the execution time. As a comparison test, we generated another large dataset whose file size is more than 600 MB. The results are plotted in Figure 4 and Figure 5.

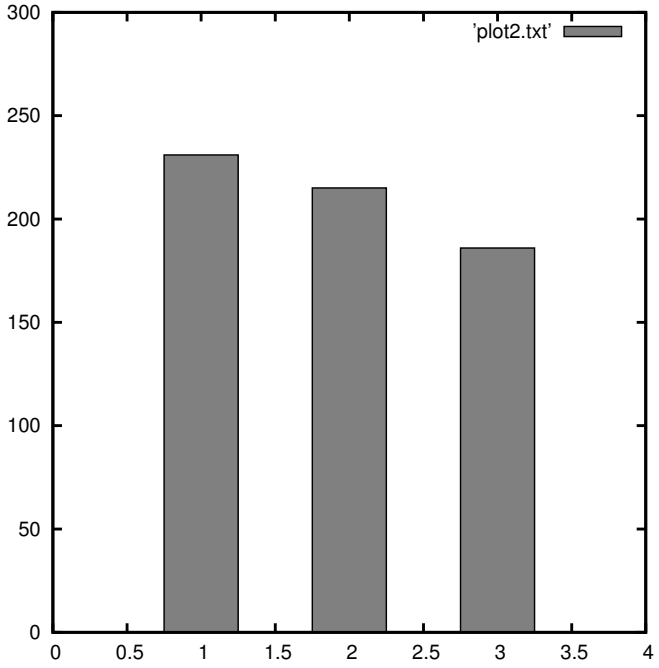


Figure 4: Execution time for different number of nodes in the 2M data

In the second experiment, we chose different time slots: half a second, 1 second and 2 seconds. Then we counted their frequency tables according to event logs. The matrices are stored in text files using the KV format. Then we calculated the correlation matrix of each frequency table and got a result matrix representing the correlation between two types of events. The results are plotted in Figure 6 and 7.

In another experiment, we applied random projection to reduce the amount of data. We tried different values for the random matrix sparsity and the reduced results are listed below:

- Number of events in simulation data: 28749
- Number of events in synthetic data: 263783
- Number of events in combined data: 292532
- Number of events when 20% off: 234086
- Number of events when 50% off: 146124

The random projection can effectively reduce the amount of data but what we care most is about the patterns in the reduced data rather than the number of events. In the next experiment, we calculated the correlation matrix as in the second experiment, and plotted the results for comparison. The Figure 8 and 9 show the results.

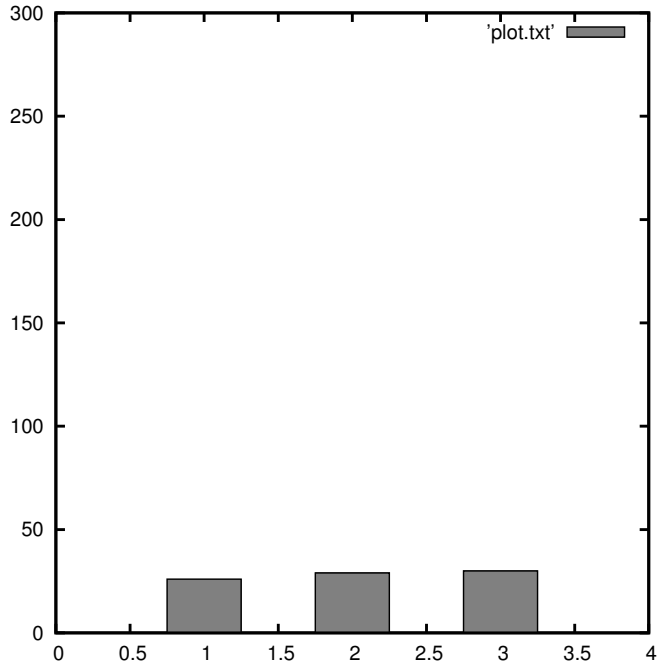


Figure 5: Execution time for different number of nodes in the 600MB data

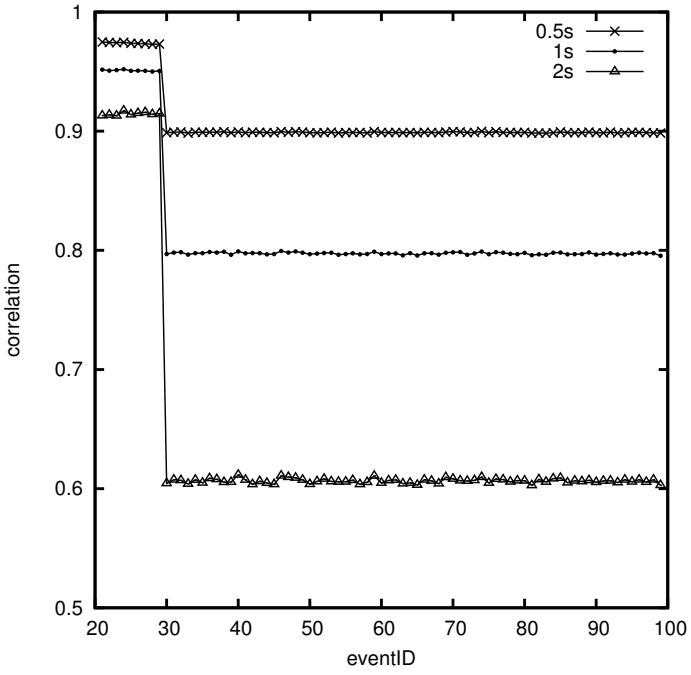


Figure 6: Correlation between Event 20 and Event 21∼99 in the combined data

5 Analysis and Evaluation

In this section, we analyze the results we got from experiments. According to the figures, we were able to infer some characteristics in the raw data and we also evaluate the effect of Hadoop and random projection.

Figure 4 shows the effect of merging and sorting on small size data and Figure 5 on large size data. When the data size is small (2MB), the execution time is not very much influenced by using more nodes. However, when the data size is large (600MB), it is obvious that the running time decreases

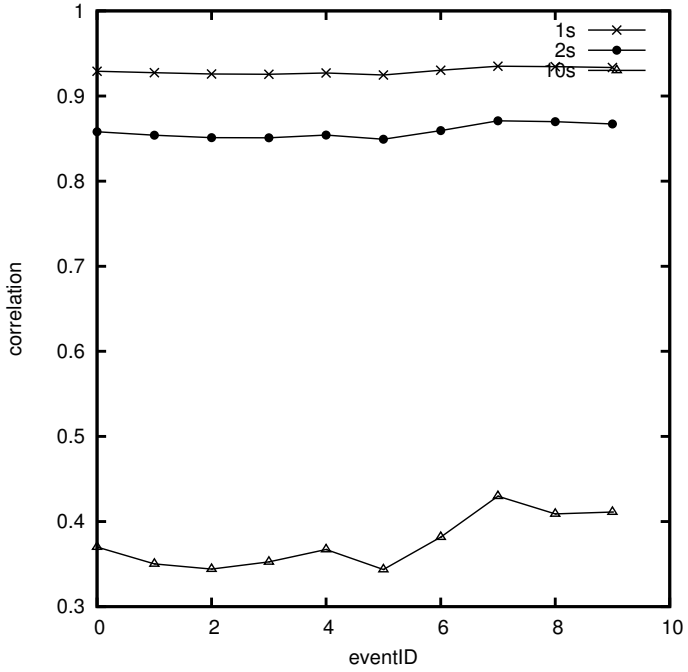


Figure 7: Correlation between Event 10 and Event 0∼9 in the combined data

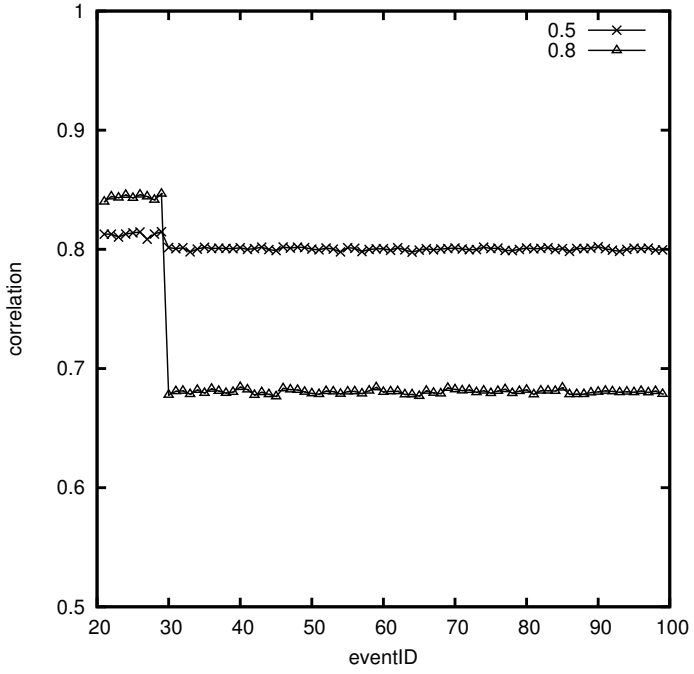


Figure 8: Correlation between Event 20 and Event 21∼99 in the reduced data

when using more nodes. If there were enough number of nodes, the running time could be very fast. The experimental result indicates that Hadoop is more suitable for large files rather than small files. In Hadoop, if the dataset is less than the default block size, it cannot be divided into smaller blocks and be distributed to several nodes and so there is no benefit from parallel computing.

The Figure 6 shows the correlation between event 20 and event 21∼99, which belong to the synthetic data set. Taking event 20 as example, it is easy to find a gap in the lines. The value of correlation descends suddenly between event 29 and event 30. According to the figure, we can infer that the event 21∼29 have

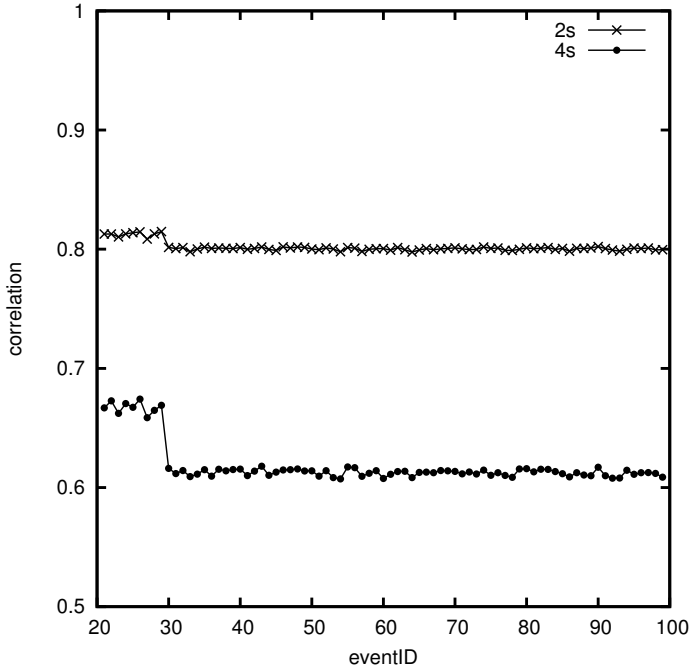


Figure 9: Correlation between Event 20 and Event 21∼99 on 50% off data

a stronger relationship with event 20. It means that the events 20∼29 are likely to occur together. As the rest of the events have lower correlation coefficient, it means that they seldom occur dependently. Because, we defined a pattern that EventIDs are from 0 to 9 when generating synthetic data, the analysis result meets our expected goals.

The Figure 7 shows the correlation between event 10 and event 0∼9. Because these events come from the simulator data, it is difficult for us to verify our analysis results directly. What we can conclude is that the correlation matrix is sensitive to time slots. Taking event 10 for example, the result is very similar when using 1-second and 2-seconds time slot. But when the time slot is 10 seconds, the difference of correlations becomes apparent. It shows that there are stronger relationships between events 7, 8, 9 and 10. In other words, event 10 is more likely to happen with event 7, 8, 9. For forecasting, if event 7, 8, 9 occur, it is very likely that event 10 will also happen.

The Figure 8 shows the results on the reduced data. The 0.5-line represents the correlations in 50% off the data, which means the reduced data is half the original data. Similarly, the 0.8-line represents the correlations in 20% off the data, which means the reduced data retains 80% of the original data. The correlation difference in 0.5-line is not easy to detect. However, we can find an evident gap in the 0.8-line. Because the blue line kept more original data, which means more information and more patterns, it is easier to find patterns.

The Figure 9 shows the results on 50% off data using 2-seconds and 4-seconds time slots. It shows that correlation difference in the blue line is more obvious than that in the red line. We assumed the frequency of patterns occurring should be less when reducing 50% off the original data, but the reduction does not affect the patterns themselves. In other words, the patterns are still in the reduced data, except that the their frequency is lower. So when we used a larger time slot, it can smooth the effect of lower frequency and we are able to get the expected result.

When we looked into the running logs in Figure 9, we found that the computational complexity of 2-seconds timeslots is double of 4-seconds timeslot. Suppose the reading and writing speed of the disk is constant, it takes half time for processing the data for the 4-seconds time slot. The main reason behind this is that, with smaller timeslots the number of rows will increase in the frequency matrix. In the case of Petabyte scale data, the improvement in processing time will be significant with optimal length timeslot.

6 Conclusion

This work presented a 'big data' analytics framework for extracting the correlative structure between the events or instances. One of the key functionality of the proposed framework is to reduce data through a refined approximation process which is based on randomization and roughly preserve the statistical relationship between the instances. The approximation process is applied to multiple copies of partitioned data first for scalable performance and then for producing more accurate approximation of the data. This entails faster and good correlation computation.

The framework is realized in the MapReduce parallel model for handling massive amount of data. The execution time decreased as more nodes were used. Removing bulk of the events through randomization and then keeping only the set of events linked through statistical or temporal dependency endowed in such scale of reduction. Future work will involve applying our algorithms on real telecom network data. The data will be hundreds of gigabytes and we will build a Hadoop cluster with more nodes.

Acknowledgements

This work was partly funded by Enterprise Ireland Innovation Partnership Programme with Ericsson Ireland under grant agreement IP/2011/0135 [Dublin City University and Ericsson, 2014].

References

[Achlioptas, 2003] Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of Computer and System Sciences*, 66(4):671–687.

[Bingham and Mannila, 2001] Bingham, E. and Mannila, H. (2001). Random projection in dimensionality reduction. In *Proc. seventh ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '01*, pages 245–250, New York, New York, USA. ACM Press.

[Butalia et al., 2008] Butalia, A., Dhore, M., and Tewani, G. (2008). Applications of Rough Sets in the Field of Data Mining. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 498–503. IEEE.

[Chen, 2010] Chen, Z. (2010). Research of Data Mining Based on Neural Network. In *2010 International Conference on E-Product E-Service and E-Entertainment*, pages 1–3. IEEE.

[Cisco, 2013] Cisco (2013). Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012 2017.

[Cloudera, 2010] Cloudera (2010). Cloudera Helps Hadoop Summit Reach New Heights.

[Dean and Ghemawat, 2008] Dean, J. and Ghemawat, S. (2008). MapReduce. *Communications of the ACM*, 51(1):107.

[Dublin City University and Ericsson, 2014] Dublin City University and Ericsson (2014). E-Stream Project.

[Ghemawat et al., 2003] Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003). The Google file system. *ACM SIGOPS Operating Systems Review*, 37(5):29.

[Li et al., 2006] Li, P., Hastie, T. J., and Church, K. W. (2006). Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '06*, page 287, New York, New York, USA. ACM Press.

[Patel et al., 2012] Patel, A. B., Birla, M., and Nair, U. (2012). Addressing big data problem using Hadoop and Map Reduce. In *2012 Nirma University International Conference on Engineering (NUiCONE)*, pages 1–5. IEEE.

[Piro et al., 2011] Piro, G., Grieco, L., Boggia, G., Capozzi, F., and Camarda, P. (2011). Simulating lte cellular systems: An open-source framework. *IEEE Transactions on Vehicular Technology*, 60(2):498–513.

[Wei Gao et al., 2011] Wei Gao, Kun Niu, Man Cui, and Qingyang Gao (2011). A data prediction algorithm based on BP neural network in telecom industry. In *2011 International Conference on Computer Science and Service System (CSSS)*, pages 9–12. IEEE.

[Xu et al., 2012] Xu, G., Xu, F., and Ma, H. (2012). Deploying and researching Hadoop in virtual machines. In *2012 IEEE International Conference on Automation and Logistics*, pages 395–399. IEEE.

[Yuan et al., 2014] Yuan, Z., Keeney, J., van der Meer, S., Hogan, G., and Muntean, G.-M. (2014). Context-aware Heterogeneous Network Performance Analysis: Test-bed Development. In *PerCom 2014 - 11th Workshop on Managing Ubiquitous Communications and Services*. IEEE.

[Zaman et al., 2014] Zaman, F., Robitzsch, S., Wu, Z., Keeney, J., van der Meer, S., and Muntean, G.-M. (2014). A heuristic correlation algorithm for data reduction through noise detection in stream-based communication management systems. In *NOMS 2014 - Data Centric Cloud and Network Management*.

An Artificial Neural Network Based Approach to Improve Building Energy Efficiency

Ravi Prakash, Robert Perry, James Mooney, Enda Fallon

Software Research Institute, Athlone Institute of Technology, Athlone, Ireland
A00214413@student.ait.ie, kelevra88@hotmail.com, jamesmooney@ait.ie, efallon@ait.ie

Abstract

Due to the ever increasing cost of energy associated with buildings it is desirable to minimize energy consumption by controlling its use in an intelligent manner. In this paper an artificial neural network based temperature control system is proposed and implemented to optimize building energy consumption. The proposed perceptron based neural network model is experimentally verified. The developed system consists of a server side implementation which interfaces to sensors and effectors using the Arduino platform. Results obtained illustrate the capability of the developed perceptron model to learn from historic behaviour in order to mediate between ambient temperature, desired temperature and power consumption. The results also highlight the potential for the use of this model in a practical building energy management system in order to minimize energy consumption.

Keywords: Artificial neural network, perceptron model, temperature control, smart buildings.

1 Introduction

In 2008 the U.S. Department of Energy (DOE) stated that among the various energy consumption sectors, residential and commercial buildings amounted to 41 per cent of the total energy consumption in the U.S. [1]. With the increasing prices of energy resources, soaring power bills are becoming a burden on each section of society, whether it's residential or industrial and hence are motivating research to find a stable and easily adaptable solution using different technologies. A solution is to use renewable energy resources such as solar panels. Another approach is energy conservation. There is therefore a need to focus on building management systems to control and manage the energy consumption in households or industries. The aim is to build a cost effective model that can predict the energy needs based on energy utilization patterns. The system should also suggest a better management technique supported with data and facts. An accurate and reliable energy prediction scheme combined with an automated energy data collection system can assist building managers identify maintenance issues and determine the best energy control strategies. An automated energy prediction system can be built on top of a mathematical prediction model consisting of several parameters.

Although rules based systems are implemented in most existing heating control systems, they are not ideal for reducing energy consumption, especially in a commercial environment. An example is where a static temperature controller checks the temperature of a system at a specific time interval and makes changes if the system is outside the parameters specified. This system could be applied where the ambient temperature exceeds the set-point temperature in a room and hence the cooling system needs to be activated. Alternatively it could be applied to a heating system which tries to increase the ambient temperature when the room falls below a certain set-point. These rules based systems, while simpler to implement from a programming point of view, cannot attain optimal efficiency levels because they result in large fluctuations in temperature about the set-point.

A variety of prediction models have been proposed in the literature including time-series models, Fourier series models, regression models, Artificial Neural Network (ANN) models and Fuzzy logic models. Each model has its own features, advantages and disadvantages, and in addition, its performance varies from one application to another. Here it is proposed to use the ANN model to conserve energy consumption as well as learning simultaneously to predict the future energy consumption of a building using the back propagation method.

Even though modern computer systems are able to outperform the human mind in numeric computations and related symbolic manipulation still many complex perceptual tasks (like face recognition and voice recognition, etc.), can be performed by the human mind effortlessly in contrast with modern computer systems. ANN tries to imitate all those abilities of human like massively parallel computing systems and also learns simultaneously with the feed forward mechanism as well as being fault tolerant. These characteristics therefore make ANN a suitable choice for the energy conservation application [2]. This paper aims to show that the proposed model effortlessly and accurately controls the energy consumption and learns with time to conserve energy more efficiently.

The paper is organized as follows. Related work is discussed in Section 2. This is followed in Section 3 by an introduction to the artificial neural network model and in particular the perceptron model. Section 4 describes the system architecture used in applying the neural network model to the temperature control system. An overview of the experimental implementation is presented is Section 5 while experimental results are discussed in Section 6. Section 7 provides conclusions and ideas for future work.

2 Related Work

There has been a large amount of research in this field due to the pertinence of the work. A number of papers present the advantages of ANN in energy systems for conservation as well as prediction. In [3], the authors present an approach for forecasting electrical energy consumption of equipment maintenance based on ANN and particle swarm optimization (PSO). The authors explain how ANN was used for mapping the relationships between the input variables and expected electrical energy consumption and how a new PSO algorithm with adaptive control parameters was used to evolve the ANN in a more accurate way.

In [4], Ismail et. al. consider a gas station scenario and illustrate how more convenient it would be for the gas distributors to have an accurate prediction system. They emphasize the desirability of a dynamic prediction model that can adapt itself to changes in the energy consumption patterns for short-term energy prediction and also show through experimental work how the sliding window method is a better approach than the accumulative training method in this scenario.

In [5] the authors recommend not using static prediction models which involve a single prediction model that does not evolve over time. This makes the model slightly rigid because when the training is complete, the model will be fixed and will not use the most recently collected data. Hence a dynamic prediction model is needed that can adapt itself with the changes in the energy consumption patterns. The authors also introduced the two types of adaptive model, i.e. sliding window and accumulative training and showed with the help of experiments the behaviour of both models, illustrating that for a short term prediction like energy consumption for a building the sliding window model is a better approach than the accumulative training model. Also the authors explain why accumulative training model might work just fine with the synthetic noise free data but is not able to do well for the real world, lower quality input data which is well taken care of by the sliding window as the author has shown with the experimental results.

In [6], the author has undertaken an extensive study of various energy applications of ANN. The work illustrates the ability of ANNs to address non-linear problems, applications such as solar heating, heating, ventilating and air conditioning (HVAC) systems, air conditioning of buses and other

problems such as forecasting and prediction. It is also shown how ANN could be advantageous for each scenario. The author also suggests that to apply ANN systems, data from both the past and current performance of the real system is required in order to apply the learning and to select suitable neural-network models.

These studies illustrate the necessity of an ANN model for prediction in building energy systems. However, research on the application of ANN models together with intelligent training algorithms to the field of energy consumption forecasting is still rare.

3 Artificial Neural Network Model

Research on ANNs has seen three periods of extensive activity. Initial pioneering work was conducted in the 1940s by McCulloch and Pitts [7, 8] and then followed in the 1960s by Rosenblatt's perceptron convergence theorem. A resurgence then occurred in the 1980s mostly due to Hopfield's model and the back-propagation learning algorithm for the multilayer perceptron model.

3.1 Artificial Neural Network Model

From the biological perspective, a neuron is a biological cell which can process information in the human mind. An average human mind (cerebral cortex) has around 10^{11} neurons where each neuron is connected with 10^3 to 10^4 other neurons. All of these interconnected neurons communicate through a very short train of pulses and even though the switching frequencies are several times slower than that of high power computers available today they are still able to perform very complex perceptual decisions like face recognition in a few milliseconds [2]. The basic factor behind this advantage of neurons over Von Neumann computers is extensive parallel programming.

Inspired by biological nervous systems, an ANN consists of an inter-connection of a number of neurons. It uses processing elements connected by links of adaptive weights to form a black box representation of systems [3] to achieve success in problems in which other rule based programming could not. The artificial neural network is well known for its capability to learn from examples and deal with non-linear and complicated problems. It is also good for tasks involving incomplete-data sets and fuzzy or incomplete information. One of the applications that neural networks perform best in is forecasting and prediction [4]. A huge number of researchers have recommended the neural network technique for short-term, midterm or long term gas consumption forecasting.

3.2 Perceptron

An arrangement of one input layer of neurons feeding forward to one output layer of neurons is known as a single layer perceptron. The perceptron, basically shows a binary relation where it maps its input *x* to an output value *f(x)* in the following way:

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where *w* is the weight applied, $w \cdot x$ computes the weighted sum, and *b* is the bias, a constant term that does not depend on any input value.

Perceptron has two main steps of operation:

a. Calculate the actual output:

$$y_j = w_j \cdot x_j \quad (2)$$

b. Update the weights:

$$w_{j+1} = w_j + (y_j \cdot \alpha), \tag{3}$$

where y_j denotes the output from the perceptron and x_j is the input at the j^{th} iteration, w_j is the weight at the j^{th} iteration, w_{j+1} is the new updated weight for the $(j+1)^{th}$ iteration and α is the learning rate, where $\{0 < \alpha \leq 1\}$. In equation 2, the actual output is calculated based on both the input and the weight applied and then in equation 3, based upon this result, the weight for the next iteration is calculated where the current weight is summed up with the product of the result from equation 3 and a constant learning rate. This illustrates the true behaviour of a feed forward network. [9,10].

4 System Architecture

As illustrated in Figure 1, in the proposed system data is taken from the temperature sensors and the temperature differential is calculated based upon the desired temperature. Following the perceptron algorithm, weights are applied and activation values generated. Larger weights increase the reactiveness of the system. Lower weights reduce the reactiveness of the system. Based on the scale of activation value (weight * input value), a specific neuron is selected for firing. The neuron which fires is based on non-overlapping activation value boundaries. The specific neuron which fires identifies a corresponding effector action on the environment, as an example turn on heater or turn on cooler.

Following a learning cycle, the effect of the action on the environment is evaluated. If the action had a beneficial effect on the environment, it reduced the differential between desired and actual temperature, synaptic weights are increased. If the action had a detrimental effect on the environment, it increased the differential between desired and actual temperature, synaptic weights are reduced. Eventually the system will reach and maintain a value approximately equal to the desired temperature value. Figure 1 illustrates the system approach.

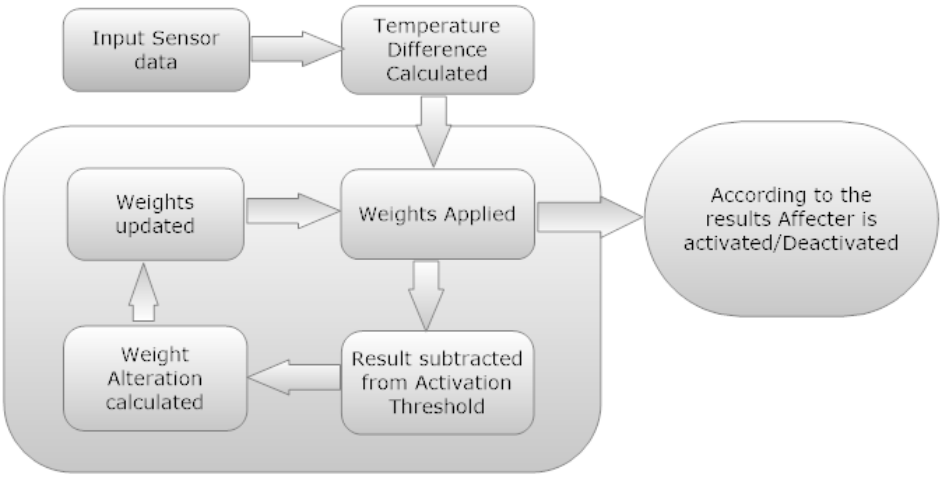


Figure 1: System approach.

5 Experimental Implementation

The experimental architecture consisted of an Arduino adapter, a heater and a server running on a laptop. The experimental configuration is illustrated in Figure 2. The server code implements the neural network. It controls the heater using the information it receives from the various temperature sensors. All elements of this setup are interconnected using the TCP/IP protocol using an existing router setup found in most homes and businesses.

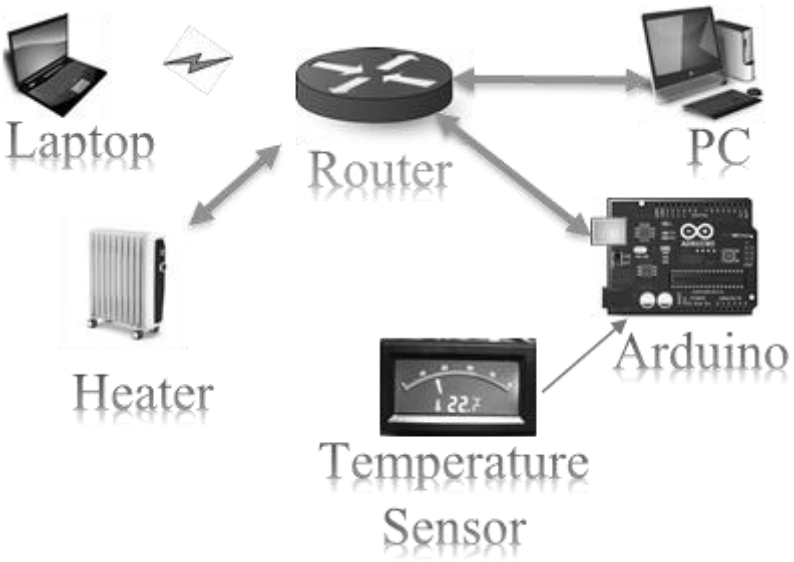


Figure 2: Experimental architecture.

The proposed architecture aims to demonstrate some advantages of controlling a heating or cooling system using an ANN approach. In the configuration a single heater is used to affect the system. The system could be easily manipulated to run many effectors (heaters/coolers/automated door opening) in a large commercial environment from just a single server setup using multiple sensors and control interfaces.

The Arduino open source electronic prototyping platform used features the Atmel Atmega series 328 microprocessor [11,12]. Communication for programming of the microcontroller and power are provided to the Arduino board via a USB connection [13]. The complete system can therefore be powered via a single 5 Volt power supply. The microcontroller has a 32 Kilobyte memory capacity enabling both room for the system code to be uploaded and also allowing future revisions and additions to the system. The board has both analogue and digital outputs and inputs. The Arduino provides the interface to the hardware infrastructure through the connection of a relay to the digital output port for control of the heater. The digital input on the board allows a digital temperature sensor to be connected directly to the board and its data displayed by means of a web server that is created using the Ethernet shield. The Arduino thereby provides the essential link between the TCP/IP network and the hardware infrastructure. This is all implemented by configuring the Arduino using the appropriate 'Arduino sketch' programming code. The client side controls all the switching functionality and hence is able to control the switching of any device on or off on a 220 Volt live line through any web browser.

An Ethernet shield is used that implements the same functionality as a conventional Ethernet module except that it has been specifically designed to interface with the Arduino platform. The shield allows the Arduino to accept connections through a Cat 5 Ethernet Cable. These requests are routed through the TCP/IP network protocol. The requests are processed by the board and passed on the Arduino platform. The two way communication feature of this shield enables having a high compatibility with existing network infrastructures. This device therefore allows the Arduino to communicate through a conventional router in an existing home or commercial setup. For example, a URL call to Http://192.168.1.200/?HeaterOn would switch the heater on.

Using a classic home network setup in conjunction with the Arduino Ethernet shield infrastructure, suitable network infrastructure was defined, that not only meets the functional requirements of the system but also incorporates encryption systems used in conventional routers to keep the network safe from outside intrusion. Using the WPA2 standard and the conventional router setup a high level of

security is ensured. The system can therefore be completely wireless, ensuring compatibility with existing systems and improving ease of installation. This will also offer the ability to connect directly should a wireless connection not be suitable, which may be the case in some existing commercial environments.

6 Experimental Results

Table 1 explains the basic functionality of the system with the use of the simple perceptron model and multi-neurons in the ANN by illustrating the results obtained for three iterations.

Table 1 Three iterations of perceptron learning process using multi-neurons.

Indoor Sensor Temperature (Degrees Celsius)	Desired Temperature (Degrees Celsius)	Indoor Temperature Difference (Degrees Celsius)	Weight	Activation Threshold			Result	Action	Learning Rate	Weight alteration
				AT1	AT2	AT3				
15	22	-7	-0.2	0.1	0.6	1.2	1.4	Heater On (Constant)	-0.01	-0.014
16	22	-6	-0.214	0.1	0.6	1.2	1.284	Heater On (Constant)	-0.01	-0.01284
17	22	-5	-0.226	0.1	0.6	1.2	1.1342	Heater On (switching off for 1 min. every 4 minutes)	-0.01	-0.011342

The indoor sensor temperature here is the temperature recorded by the sensor in the room where the experiment is undertaken. It is required to maintain the desired temperature (22^o C). In order to maintain the temperature it calculates the indoor temperature difference $(Ti)_j$. The weight, $w_j(t)$, is chosen to be -0.2 arbitrarily at the start and should be corrected by the system as learning proceeds. There are three activation thresholds which determine which neuron should be fired and when, depending on the value of the result of the multiplication of the temperature difference by the weight, as stated in 4:

$$result = (Ti)_j * w_j(t). \quad (4)$$

For the first iteration in Table 1, with a room temperature of 15°C and a desired temperature of 22°C the system detects a temperature difference of 7°C and hence generates a result of 1.4 by multiplying the difference with the assumed weight, -0.2. As the result is greater than the highest activation threshold (AT3) of 1.2, the specific neuron is fired to tell the system to carry out the specific action for that neuron, which in this case is to turn on the heater in a constant mode (or full heating mode). However for the second iteration the weight has been updated by the exact weight alteration value that was calculated in the first iteration, i.e. by -0.014, where

$$weight\_alt = result_j * learning\_rate_j \quad (5)$$

It is also clear from the second iteration that the result is still above AT3, so the system releases the same decision as the previous decision and updates the weight for the next cycle. In the third cycle, the result dips below AT3 but above AT2, and hence the neuron specific to this range is fired. The action desired in this case is to change the heater to an alternative switching mode where it will be switched to on for 4 minutes and then will go to sleep for 1 minute. Similarly for the result below AT2 and above AT1, the heater will be changed to the mode where it is switched on for a minute and then switched off for a minute alternately until the result falls below AT1 and then the heater is switched off until the result crosses any of the thresholds again. As a result the frequency of the heater's switching will increase and the heater will be shut down more frequently and hence the consumption of energy will be greatly saved.

Figure 3 illustrates the difference in desired and actual temperature. 0 on the y-axis represents the desired temperature of 22 degrees. The initial incline of the plot represents the rise in temperature in the room after the heater has been turned on. The first decline indicates where the heater has been turned off. Figure 3 illustrates an undershoot/overshoot phase until eventually the desired and actual temperatures correlate. As the system learns, shorter bursts of heating or cooling are required. These shorter bursts reduce energy consumption. This is an improvement over a conventional rule based system.

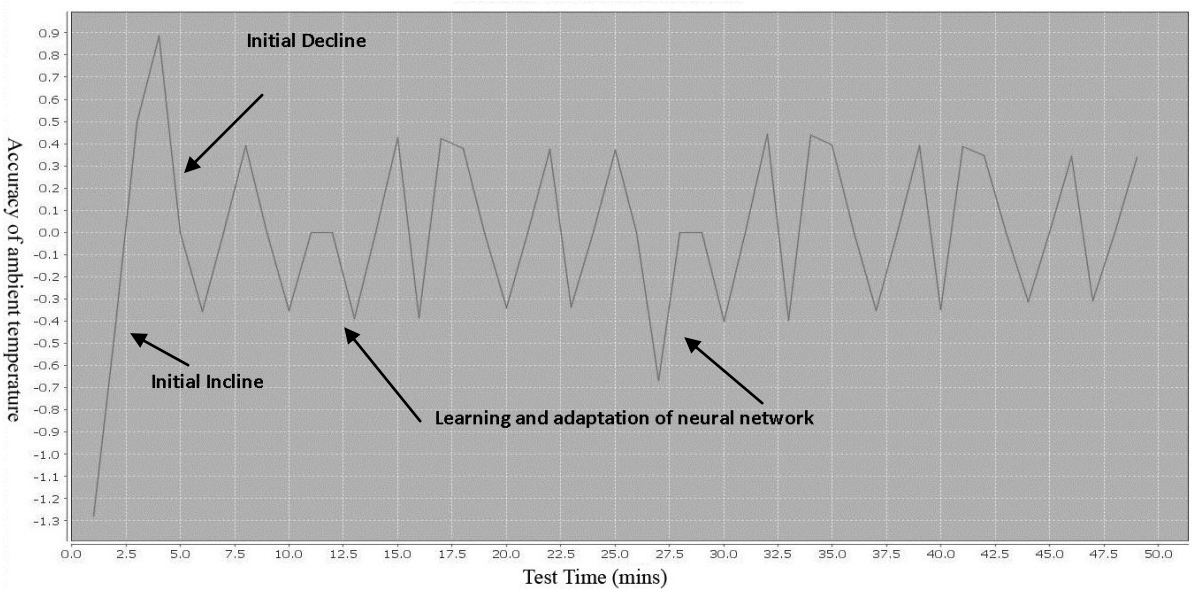


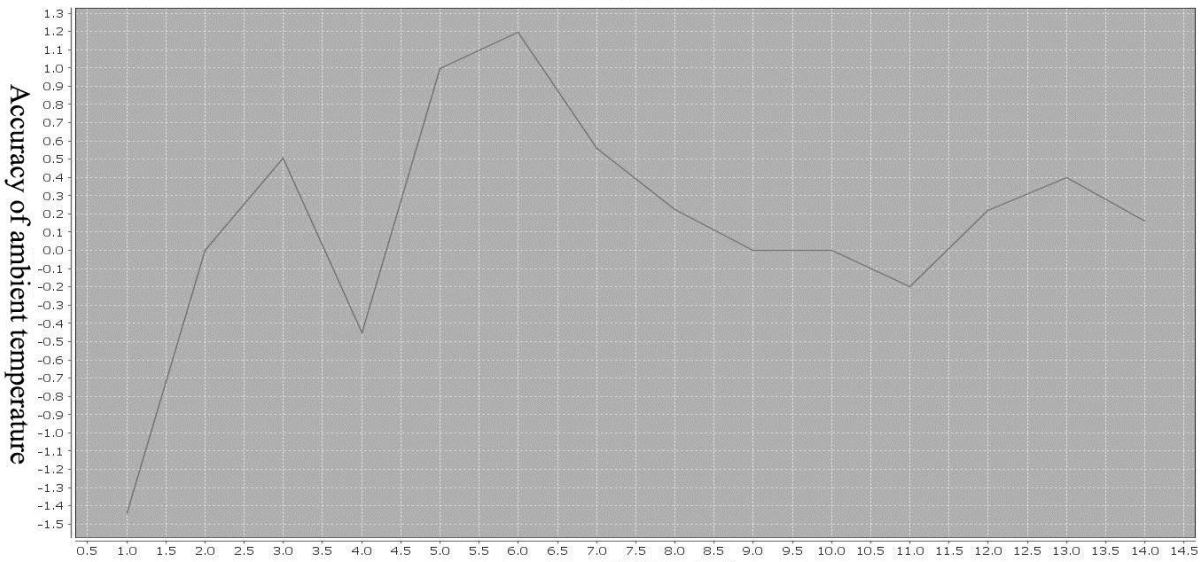
Figure 3: Neural network results illustrating learning and adaptation.

Figure 4 illustrates the learning process outlined in Figure 3 in finer granularity. Comparing Figure 4 and Figure 3, illustrates the dependence of the system on the ambient and set point temperatures. Unlike a rule system which has preconfigured decision logic, which does not adapt to changes in the system deployed environment, the ANN approach can gauge performance and learn by dynamically altering synaptic weights. As an example, if alterations are made to the construction of the building, in which the system is deployed (windows added, partitions inserted) the ANN approach will adapt the length of time the effector (heater/cooler) needs to be turned on by reducing or increasing synaptic weights. The efficiencies introduced by an ANN approach have the potential to reduce energy consumption and related financial costs.

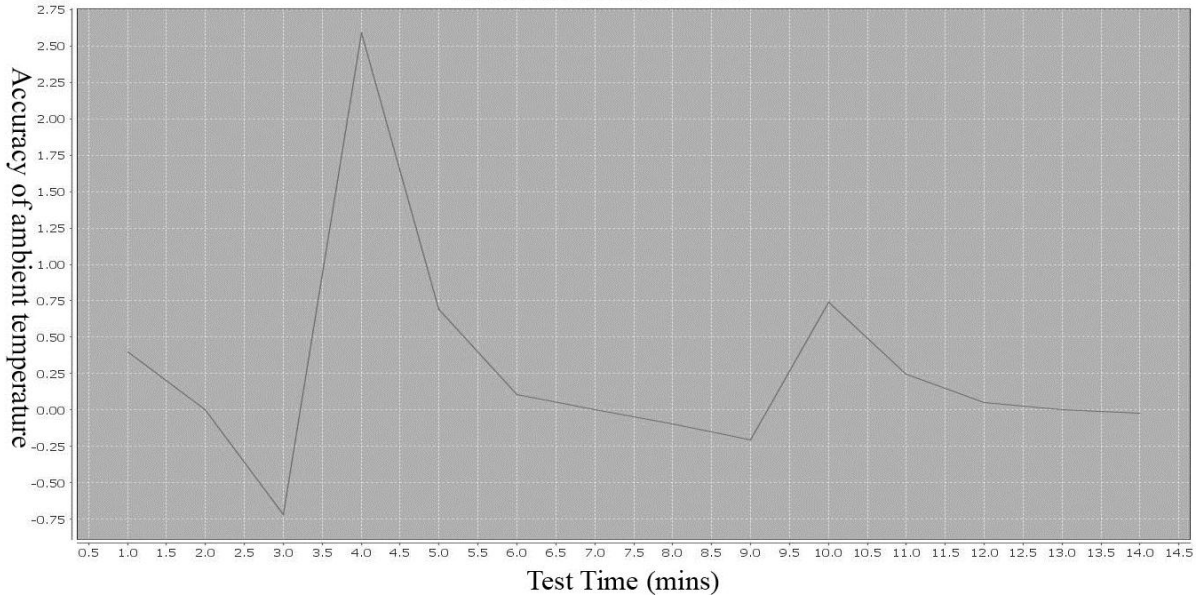
The results illustrate that the system learns how and when to make decisions on its output and should conditions change for any reason, the system will adjust itself accordingly levelling out the output until it is once again optimised.

7 Conclusions and Future Work

As a result of increasing energy costs it is desirable to optimize energy consumption. Traditionally building management systems have been developed using rules based systems. Such approaches are limited as they do not (a) adapt to changes in the environment in which they are deployed (b) learn from historic performance. In this paper an ANN based temperature control system is proposed and implemented to optimize building energy consumption. The proposed perceptron based neural network model is experimentally verified. The developed system consists of a server side implementation which interfaces to sensors and effectors using the Arduino platform. An ANN approach such as that suggested in this paper has advantages over traditional rules based building management systems. As an example, if alterations are made to the construction of the building, in which the management system is deployed (windows added, partitions inserted) the ANN approach will adapt to changed environmental conditions. The ANN will alter the length of time the effector



(a) Initial incline



(b) Initial decline

Figure 4: Neural network temperature sensor results for different input metrics.

(heater/cooler) needs to be turned on by reducing or increasing synaptic weights. The efficiencies introduced by an ANN approach have the potential to reduce energy consumption and related financial costs. Results obtained illustrate the capability of the developed perceptron model to learn from historic behaviour in order to mediate between ambient temperature, desired temperature and power consumption.

Future work will focus on the addition of additional sensory inputs to the ANN model. On-going work focuses on the quantification of the energy saving.

References

[1] N. Fumo, P. Mago and R. Luck, "Methodology to estimate building energy consumption using EnergyPlus Benchmark Models," Energy and Buildings, vol. 42, no. 12, pp. 2331-2337, 2010.

[2] A. K. Jain, J. Mao and K. M. Mohiuddin, "Artificial neural networks: A tutorial," Computer - Special issue: neural computing: companion issue to Spring 1996 IEEE Computational Science & Engineering, vol. 29, no. 3, pp. 31-44, 1996.

[3] X. Jiang, H. Ling. J. Yan, B. Li and Z. Li, "Forecasting electrical energy consumption of equipment maintenance using neural network and particle swarm optimization," Mathematical Problems in Engineering, 2013.

[4] M. J. Ismail, R. Ibrahim and I. Ismail, "Adaptive neural network prediction model for energy consumption," Computer Research and Development (ICCRD), 3rd International Conference, vol. 4, pp. 109-113, 2011.

[5] "J. Yang, H. Rivard and R. Zmeureanu, " Building energy prediction with adaptive artificial neural networks, pp. 15-18, Ninth International IBPSA Conference,2005.

[6] S. A. Kalogirou, "Applications of artificial neural-networks for energy systems," Applied Energy, vol. 67, pp. 17 - 35, 2000.

[7] I. C. Baianu, "A logical model of genetic activities in lukasiewicz algebras: The non-linear theory," Bulletin of Mathematical Biology, vol. 39, pp. 249 - 258, 1977.

[8] K. Aizawa, "Warren McCullochs Turn to Cybernetics: What Walter Pitts Contributed.," Interdisciplinary Science Reviews, vol. 37, pp. 206 - 217, 2012.

[9] A. Wasilewska. [Online]. Available: www.cs.sunysb.edu/~cse634/ch6NN.pdf. [Accessed 05 March 2014].

[10] D. M. Humphrys, "Single Layer Neural Networks (Perceptron)," [Online]. Available: http://computing.dcu.ie/~humphrys/Notes/Neural/single.neural.html. [Accessed 05 March 2014].

[11] S. F. Barrett, *Arduino Microcontroller Processing for Everyone!* 3rd ed., Morgan & Claypool, 2013.

[12] Atmel Corporation. [Online]. Available: http://www.atmel.com/devices/atmega328.aspx. [Accessed 05 March 2014].

[13] Arduino Duemilanove. [Online]. Available: http://arduino.cc/en/Main/arduinoBoard Duemilanove#. UxORTfl_srY. [Accessed 05 March 2014].

Session 3

Technical Poster Session

PHANS - A Probability Based Handover Algorithm for Network Selection Using End User Predicted Movement

Niall Maher, Shane Banks, Enda Fallon

Software Research Institute,

AIT Athlone Institute of Technology,

Dublin Road, Athlone, Co Westmeath.

Email: niallmaher@research.ait.ie, sbanks@ait.ie, efallon@ait.ie

Abstract

Traditional mechanisms trigger network selection based on dynamic performance metrics such as Received Signal Strength (RSS), delay, loss etc. Such approaches do not consider how the predictability of end user movement can be used to influence optimal handover selection. For systems with repeated movement such approaches are limited as they do not consider how the predictable nature of mobility can be used to influence network selection. Student groups in college follow predictable routes between classes. Public service vehicles such as busses and trains typically operate in preconfigured routes which are repeated at routine intervals. Traditional network selection mechanisms do not exploit this performance predictability. This work proposes PHANS – A Probability Based Handover Algorithm for Network Selection Using End User Predicted Movement in order to weigh the relative importance of both dynamic performance metrics together with predictable movement patterns in order to optimise network selection.

Keywords: Network handover, Mobility, Probability

1. Introduction:

In principle, each mobile terminal (node) is, at all times connected to a network and within range of at least one network access point on that network [1].The area serviced by a Base Station (BS) is identified as its cell. As a mobile node moves it will handover its connection from one cell on a network to another cell. Most network selection approaches use an evaluation of network performance to determine when handover should take place [2][4][5].Traditional handover initiations are based on RSS [4][5]. Such approaches have evolved by proactively predicting RSS values, though still making use of static handover triggering thresholds [6][7][8]. Other criteria that are now taken into account in handover initiations include; bandwidth, latency, link quality and Quality of Service (QoS) [3]. Such network selection approaches are limited as they do not consider how the predictable nature of mobility can be used to influence network selection. We propose to weigh the relative importance of the dynamic performance metrics together with predictable movement patterns in order to optimise network selection. The dynamic selection will be weighed against the probability selection and from this the best path will be selected in order to optimise performance.

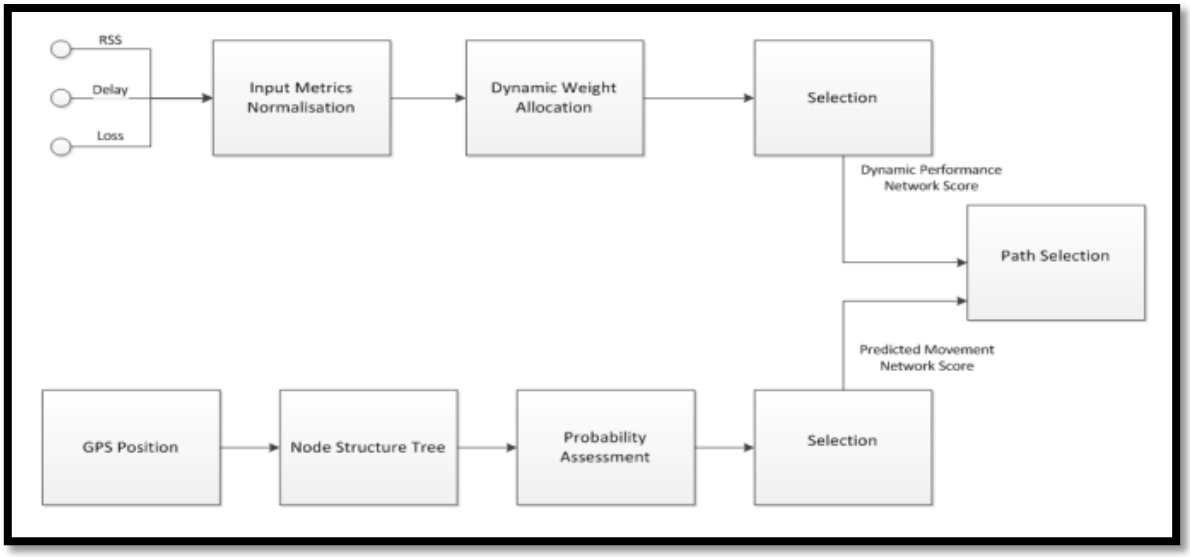
2. Related Work

Traditionally, handover algorithms utilize static thresholds that were applied to different metrics such as Delay, Loss and RSS to determine network handover. These approaches are still being used with greater accuracy by proactively predicting the RSS values, though it still makes use of the static thresholds. While RSS is an important performance parameter, used alone it does not provide an accurate view of the dynamic status of a link. That is why handover approaches have considered

multiple performance parameters which include RSS, delay and loss rate. Some solutions also consider metrics related to the content delivery quality as experienced by end users [9]. Such approaches however are performance limited as they apply static performance thresholds, which when exceeded, trigger handover. Applying a static performance threshold makes assumptions regarding the status of a network. Previous work undertaken [10] illustrates that it is beneficial for the handover management algorithm to probe network performance and dynamically alter thresholds through synaptic weights.

3. Proposed Approach

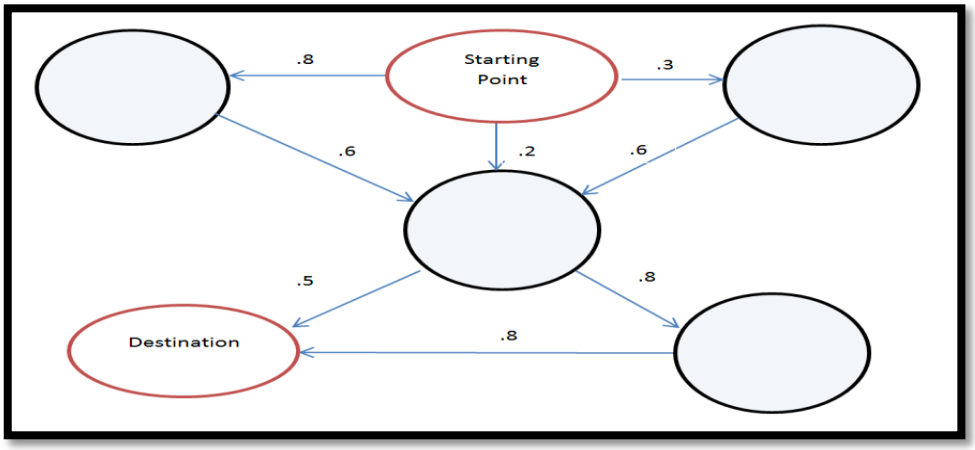
Figure 1 illustrates the major components of the PHANS algorithm.



Dynamic Network Performance Evaluation PHANS takes as input dynamic network performance metrics; RSS, Delay, and Loss. These metrics are parsed and normalised on a scale of 1 (poor performance) to 100 (excellent performance). Following performance normalisation, dynamic weight allocation provides a weight reflecting the relative importance of each metric. The dynamic weight is then multiplied by the normalised inputs and summed to indicate dynamic network performance.

Probability of Movement Assessment In the PHANS approach a nodal structure is dynamically created representing the Access Points (AP) topology in a given area. The topology is constructed based on previous mobility recorded by the GPS module on the device. Using historical data, PHANS overlays the probability of inter AP movement to the topology. The node with the highest score is suggested by the probability assessment module as the candidate AP.

Figure 2: Proximity Nodes Structured Tree



Candidate Path Selection Within the PHANS algorithm, the path selection subcomponent takes two inputs (a) a candidate AP selection based on dynamic performance (b) a candidate AP selection based on predicted movement. If these selections conflict, PHANS algorithm will weigh the previous performance of each selection when determining the current optimal path selection.

4. Results

4.1 Experimental Evaluation of Characteristics

In order to determine network performance, a series of experimental tests were undertaken at the Engineering building at Athlone Institute of Technology. The experimental tests used Linksys wireless broadband routers which were placed at specific locations to mimic the movement of students from one class to another. These tests were undertaken by physically walking a set route passing into and out of the access point coverage recording delay, loss and RSS. This data was used as input to the NS2 simulation model described below.

Figure 3: Illustrates the recorded AP RSS for the 5 APs in the configuration.

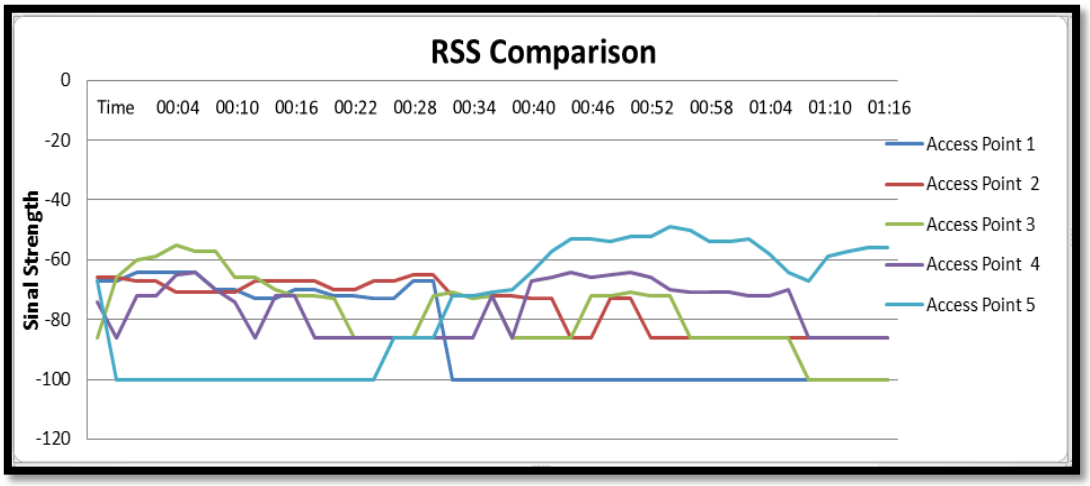


Figure 3 illustrates that at the 2 second mark the RSS indicated that AP3 was the best available AP. Using traditional handover the decision would be made to connect to AP3, but in contrast the PHANS algorithm would connect to AP2. Such an approach would avoid a potentially spurious handover to AP3. The RSS signature from Figure 4 indicates the end used is walking in the direction of access point 2 and away from AP1 and AP3. Therefore using predicted movement to influence the decision of the handover and optimise the possible throughput as seen in Figure 4.

4.2 Simulated Evaluation of the PHANS Algorithm

Figure 4 illustrates the throughput performance of the PHANS algorithm against the Stream Control Transmission Protocol (SCTP) mobility protocol. The SCTP configuration was undertaken with Path Max Retransmission (PMR) values ranging from 0 (aggressive handover) to 2 (moderately aggressive handover). The default PMR value is 5. This configuration is seen as passive for wireless mobility scenarios.

Figure 5: Throughput Performance of PHANS and the SCTP Mobility Protocol

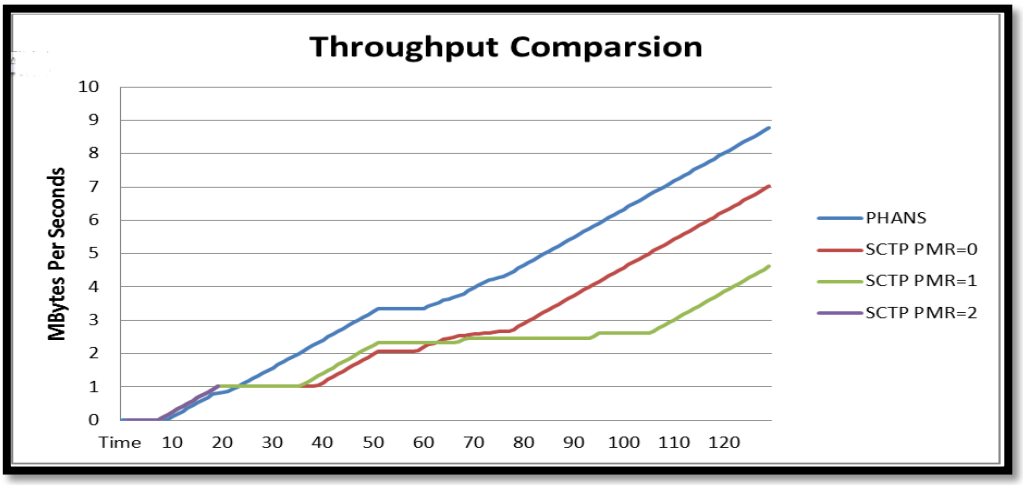


Figure 4 illustrates that PHANS transmits 83.7 MBytes during the duration of the test. SCTP PMR=0 was the most effective SCTP strategy transmitting 66.9 MBytes.

Conclusion:

In this paper we have proposed a predictability based handover using PHANS which allows us to optimize handover between end to end users, taking into account predicted movement of the end user to enhance throughput capability. Results illustrates that our predictability based handover has improved the performance by approximately 20% of the end to end user connection.

References:

[1] Nasser, N.; Hasswa, A.; Hassanein, H., "Handoffs in fourth generation heterogeneous networks," *Communications Magazine, IEEE* , vol.44, no.10, pp.96,103, Oct. 2006

[2] Stewart, R.; Metz, C., "SCTP: new transport protocol for TCP/IP," *Internet Computing, IEEE* , vol.5, no.6, pp.64,69, Nov/Dec 2001

[3] Kassar, Meriem, Brigitte Kervella, and Guy Pujolle. "An overview of vertical handover decision strategies in heterogeneous wireless networks." Computer Communications 31, no. 10 (2008): 2607-2620.

[4] NIST, "The Network Simulator NS-2 NIST Add-on – Neighbour Discovery", January 2007

[5] Li, Mingxin, Shanzhi Chen, and Dongliang Xie. "A multi-step vertical handoff mechanism for cellular multi-hop networks." In *Proceedings of the 2nd ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks*, pp. 119-123. ACM, 2007.

[6] Ben-Jye Chang; Jun-Fu Chen, "Cross-Layer-Based Adaptive Vertical Handoff With Predictive RSS in Heterogeneous Wireless Networks," *Vehicular Technology, IEEE Transactions on* , vol.57, no.6, pp.3679,3692, Nov. 2008

[7] Lyu-Han Chen; Gen-Huey Chen; Ming-Hui Jin; Wu, E.H.-K., "A Novel RSS-Based Indoor Positioning Algorithm Using Mobility Prediction," *Parallel Processing Workshops (ICPPW), 2010 39th International Conference on* , vol., no., pp.549,553, 13-16 Sept. 2010

[8] Capela, N.; Soares, J.; Neves, P.; Sargento, S., "An Architecture for Optimized Inter-Technology Handovers: Experimental Study," *Communications (ICC), 2011 IEEE International Conference on* , vol., no., pp.1,6, 5-9 June 2011
doi: 10.1109/icc.2011.5962683

[9] S. Fu and M. Atiquzzaman. SIGMA: A Transport Layer Handover Protocol for Mobile Terrestrial and Space Networks. In J. Ascenso, L. Vasiu, C. Belo, and M. Saramago, editors, Invited book chapter in e-Business and Telecommunication Networks, pages 41–52. Springer, 2006

[10] E. Fallon, L. Murphy, J. Murphy, C. Ma "TRAWL – A Traffic Route Adapted Weighted Learning Algorithm" Proc. 9th International Conference on Wired/Wireless Internet Communications (WWIC 2011) Vilanova i la Geltrú, Catalonia, Spain, , Springer Lecture Notes in Computer Science 6649, pp. 1--14

On the Use of K-NN in Intrusion Detection for Industrial Control Systems

Pedro Silva, Michael Schukat

Discipline of IT
National University of Ireland Galway
Galway, Ireland
gygabyte@gmail.com, michael.schukat@nuigalway.ie

Abstract

Intrusion detection systems (IDS) monitor network or system activities for malicious activities or policy violations. As passive and non-intrusive safeguards they are particularly useful in mission-critical networks like industrial control systems (ICS). Such systems are particularly vulnerable to malicious attacks, which have seen a significant increase in recent years. However, IDS in ICS require different approaches to intrusion detection, which go beyond conventional blacklisting / whitelisting approaches. This paper examines a new technique, which is based on using the K-Nearest Neighbour scoring algorithm to discover periodic patterns in ICS network traffic.

Keywords: Intrusion Detection, Industrial Control System, K-Nearest Neighbour, Pattern Whitelisting

1 Introduction

Industrial Control Systems (ICS) are widely deployed in large and (mission-) critical infrastructures including power grids and water / gas distribution systems. The security of Industrial Control System has been a major concern over recent years, as failures in such systems may put at risk the health and safety of entire communities [8]. In recent years we have seen a noteworthy number of incidents, including Stuxnet which is considered the most sophisticated cyber-attack targeting an Industrial Control System [4].

Existing Industrial Control Systems deployed around the world have not been designed with security in mind. These systems do not implement any secure communications and have been deployed decades ago. At that time it wasn't expected to have these systems connected to other networks as local office networks or directly to the Internet. However, as times evolved ICS systems are becoming accessible from the outside world but no security measures are applied which makes them highly vulnerable to attacks [5].

Another problem is the issue of cyber-fragility, e.g. the non-ability of some complex ICS to compensate for small changes in their operational environment. For example, it has been reported that small changes in communication latency times as a result of software / firmware upgrades can bring down an entire control system [6].

In this context, Intrusion Detection Systems (IDS) can play an important role to make ICS more secure. A typical IDS for networks usually does not require any change in the control system. An IDS is a passive component that captures network traffic to be analysed. The result of that analysis will determine if an intrusion has been detected.

Typically the network traffic of an ICS is highly periodic which usually leads to highly predictable and deterministic network traffic patterns. The work presented in this paper leverages this periodicity to determine anomalies in network traffic using k-NN algorithm.

The paper is organized as follows: In Section 2, we review related work in Intrusion Detection Systems for Control Systems. In Section 3, we explain our intrusion detection approach. In Section 4, we present our conclusions and discuss future work.

2 Related Work

Barbosa et al. [1] presented an approach based on the traffic periodicity of SCADA networks. They realized there are some SCADA networks that its normal behaviour consists only of periodic bursts of packets generated by polling mechanisms. It is assumed that a considerable number of intrusions attempts or anomalies may disrupt the traffic periodicity. There are attacks such as *Denial of Service* that can be easily detected as the attacker will send a large volume of packets in a short period of time. Attacks that do not generate large volume and only need a few packets might not be detected. However, the attack can be detected by its effect in the SCADA system. A *Buffer overflow attack* optimally only requires one malicious packet, but often the targeted process crashes. This event will cause interference in the normal traffic patterns which will be detected. The proposed approach consists of four modules. The first module captures network traffic in a passive mode. This module filters out all packets that are not relevant to SCADA processes. The second module will construct the network flows using the server-side transport port to isolate traffic. Additional attributes might be needed if the service is receiving requests from more than one client. The flows are created by sampling a fixed interval P. The sample frequency *SF* is defined as $SF = 1 / P$. For each sample the number of captured packets is stored. Before initiate detection, the *periodicity learning* module will extract the frequency of the periodic bursts and their size. This step is performed offline and results validated by operators. As detected flows are matched to known *periodic burst* it will be analysed in order to find anomalies. An alarm is raised when an anomaly is detected. A Proof of Concept is explained to demonstrate the feasibility of the approach. A spectrogram-based anomaly detection is implemented and data capture is emulated by using previously collected data. The results show that the use of a spectrogram is feasible but as the analysis is performed manually, the entire system is not viable without being automated. Also, it is concluded the algorithm is too sensitive to noise such as networks delays.

Cheung et al. [2] proposed a model-based approach to detect intrusions in SCADA networks. These models specify the expected behaviour of the system and the IDS generates an alarm when deviations are found. Several models approaches are discussed. Protocol-level models are introduced to find deviations in specific fields of Modbus protocol such as function codes. Dependent fields models are also discussed, i.e., a value of a field may depend of the value of another field. These models have been implemented using Snort which is an open-source network detection system. It is also discussed more complex models that possible involves multiple Modbus requests and responses. In this case, the authors preferred to write a custom IDS due to Snort limitations and complex architecture. However, it is stated that this IDS will be based on formal models. Prior knowledge about the deployed ICS infrastructure will be required to these approaches be effective which in some cases is not accurate.

3 Our Approach

We assume that ICS network traffic is highly periodic. This observation is based on the nature of such systems where usually Programmable Logic Controllers (PLC) are used to control automated tasks which, in most cases, are executed periodically.

3.1 K-Nearest Neighbour

Case-Based Reasoning (CBR) is the process to solve novel problems (case) with solutions of similar solved problems (case base) [7]. This process is accomplished by performing the following functions: *Retrieve*, *Reuse*, *Revise* and *Retain*. K-Nearest Neighbour is a scoring algorithm commonly used in CBR for retrieving similar cases.

The K-Nearest Neighbour (k-NN), in a CBR context, allows to query a case base in order do find cases which are the most similar to the issued query [7]. In our context, each case is a packet captured from the network. The similarity of each packet in the case base to the queried packet is calculated

through a *similarity function* [7] and is represented as a real number in *[0, 1]*. The retrieved cases can be the *k* most similar cases or a threshold can be defined and the cases which surpass this threshold are considered similar. A case is often represented as a set of attribute-value pairs [7]. To each attribute is defined a *local* similarity function $lsim(a_i,b_i)$ where a_i is an attribute of a case. A weight can be defined for each attribute and it is defined by the function $w(a_i)$ for a given attribute a_i. The weight allows changing the importance of an attribute. A global similarity is calculated weighted as an average of the local similarities.

3.2 Intrusion Detection Algorithm

Our approach is based on ICS network traffic being highly periodic. Due to this fact, these networks have a tendency to be predictable and deterministic [1]. This also means that it expected to have very similar packets crossing the network. We believe that most of the times these similar packets will be sent based in a pattern which might be possible to identify by using the k-nearest neighbour (k-NN) method. The k-NN method searches by performing queries to a given set of objects and returns a set with a score for each object. This paper proposes the application of this method to find similar packets crossing a network and with that identify patterns. As mentioned earlier each packet needs to be represented as a set of attributes. A similarity function and weight is also associated (see Table 1).

Attribute	Sim. function	Weight
Source MAC Address	Equal	1.0
Dest. MAC Address	Equal	1.0
Source IP Address	Equal	1.0
Dest. IP Address	Equal	1.0
Source TCP Port	Equal / ModbusTCPPortEqual	1.0
Dest. TCP Port	Equal / ModbusTCPPortEqual	1.0
Modbus ProtoId	Equal	1.0
Modbus Length	Equal	1.0
Modbus Unit ID	Equal	1.0
Modubus Function Code	Equal	1.0

Table 1: Attributes of an packet and correspondent similarity function and weight

The *Equal* function is used on all attributes and returns 1.0 if values are equal or 0 if not. However, when calculating the similarity of two given packets, different similarity functions of TCP Port attributes are used, depending on the context. The *ModbusTCPPortEqual* function calls *Equal* function if the port is equal to 502 in one of the packets and returns the result or just returns 1.0 if the port is not 502 in either packet. In a TCP connection, when a client connects to a server, a random port number is associated to the client. This port is the Source TCP Port. If this connection, for some reason, terminates and a new one is established, the Source TCP Port most likely will be different. Therefore, when finding similar packets in the whitelist, the *ModbusTCPPortEqual* function is used to address this fact or otherwise actual similar packets may not be considered similar. When searching for similar packets in other lists such as list of live captured packets, *Equal* function is used. The approach proposed by this paper works as follows. First, the system needs to capture normal traffic which will be the whitelist. There can be more than one whitelist. From the whitelist we also build through manual observation a set of models which allows the algorithm to perform a model-based detection when needed. These models don't need a formal specification of the ICS which we believe is an advantage. The live capture is continuous but it is not processed as a stream. A time window *t ms* needs to be set and the system delivers to the intrusion detection algorithm a list *l* with packets per iteration *t*. The number of packets may differ between iterations since these blocks are delimited by time. The intrusion detection algorithm performs the following operations per each packet p_i in *t*. First, tries to find if a packet similar to p_i has been processed. It does this by running a k-NN query to a list of packets previously processed. If it can find similar packets, then it doesn't process p_i and jumps to p_{i+1}. If no similar packet is found, then it adds p_i to that list. Next, the algorithm will perform a k-NN query to list *l* and return a list sim_l with all similar packets to p_i. Next step is performing the same k-NN query to the whitelist and get a list sim_{wl} with all similar packets to p_i. If sim_{wl} is empty, then we can infer that the p_i is not known and will be considered as an anomaly. The algorithm signals the

anomaly and jumps to p_{i+1}. If sim_{wl} is not empty, the algorithm now has two lists of similar packets: sim_{l} and sim_{wl}. Each packet p_j in these lists has a timestamp t generated when p_j was captured. Next step is calculating using function $CalcSeq(sim_x)$ the lists $diff_{l}$ and $diff_{wl}$, from, respectively, sim_{l} and sim_{wl}, which will contain the timestamp difference between two consecutive packets p_j and p_{j-1}. Therefore, given funtion $t(p_j)$ that returns the timestamp of a packet p_j we have the function $CalcSeq(sim_x) = \{t(p_1) - t(p_0), \ldots t(p_j) - t(p_{j-1}) \}$, for each p in sim_x. Hence, $diff_{l} = CalcSeq(sim_{l})$ and $diff_{wl} = CalcSeq(sim_{wl})$. Next step is trying to find a subset in $diff_{wl}$ that matches $diff_{l}$. Basically, the algorithm is trying to determine if the pattern observed in $diff_{l}$ can be found in the $diff_{wl}$. The timestamps do not have to be exactly the same. We allow the definition of a margin in *ms* which tries to mitigate some slight differences in the pattern caused by other factors such as network delays and clock imprecisions. If no match can be found, then the algorithm will try to find a match in the models defined. A model is defined by a list of possible time ranges for a particular packet. Each item in the list has a high range and a low range values. A set of these models for a particular packet defines a connection model. A packet can be associated to more than one connection model. We refer to a set of connection models as an ICS model. The algorithm looks for a sequence of models that matches the time difference list of the live capture.

Software clock imprecisions of modern Operating Systems, such as Linux, may impact the ordering of packets crossing the network over time [3]. This is based on the assumption that packets might be sent within a multithreaded environment where multiple threads share connections and send different packets among them. Usually *systems calls* such as *select(2)* are used by developers to control network operations which creates the frequency patterns. Our algorithm using k-NN determines this frequency pattern for each packet and, therefore, circumvents the possible out of order of packets issue caused by clock imprecisions. However, if different *threads* send similar packets, then these frequency patterns can also change over time and k-NN will not be able to solve this issue.

4 Conclusion

This paper presents a new approach, based on CBR and k-NN, to perform intrusion detection in the context of ICS, which typically shows highly periodic network traffic patterns. As far as we are aware of there is no other work where an inspection is performed packet by packet, taken into account patterns including timestamp differences between packets. In future work, results will be gathered running against multiple datasets.

5 References

[1] Barbosa, R. R. R., Sadre, R., & Pras, A. (2012, September). Towards periodicity based anomaly detection in SCADA networks. In *Emerging Technologies & Factory Automation (ETFA), 2012 IEEE 17th Conference on*(pp. 1-4). IEEE.

[2] Cheung, S., Dutertre, B., Fong, M., Lindqvist, U., Skinner, K., & Valdes, A. (2007, January). Using model-based intrusion detection for SCADA networks. In *Proceedings of the SCADA security scientific symposium* (pp. 1-12).

[3] Etsion, Y., Tsafrir, D., & Feitelson, D. G. (2003, June). Effects of clock resolution on the scheduling of interactive and soft real-time processes. In *ACM SIGMETRICS Performance Evaluation Review* (Vol. 31, No. 1, pp. 172-183). ACM.

[4] http://www.symantec.com/connect/blogs/stuxnet-using-three-additional-zero-day-vulnerabilities (accessed on 2014, March)

[5] Knapp, E. (2011), Industrial Network Security: Securing Critical Infrastructure Networks for Smart Grid, SCADA, and Other Industrial Control Systems, *Syngress*, ISBN 1597496456

[6] Langner, R. (2011). Robust Control System Networks: How to Achieve Reliable Control After Stuxnet. *Momentum Press*, ISBN 1606503006.

[7] Lopez De Mantaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., ... & Watson, I. (2005). Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review*, *20*(03), 215-240.

[8] Macaulay, T, Singer, B.L. (2012), Cybersecurity for Industrial Control Systems, *Auerbach Publications*, ISBN 1439801967

Has Mobile Technology Affected our Critical Thinking Skills?

David Williams [1], Cristina Hava Muntean [2]

[1] National College of Ireland, School of Computing, IFSC, Mayor Street, Dublin, Ireland
David.Williams1@student.ncirl.ie

[2] National College of Ireland, School of Computing, IFSC, Mayor Street, Dublin, Ireland
Cristina.Muntean@ncirl.ie

Abstract

We find ourselves in an age where there are over 1 billion estimated smartphones used globally, and with the advent of modern mobile technologies, information is seldom more than a few clicks away. This fact leads to the question: with information from others at our fingertips, why think for ourselves? The purpose of this study is to create a framework to improve the visibility of relevant and reliable information though existing resources. The research presented in this paper investigates how information is being sourced by students and proposes a solution that improves learner critical thinking in the context of searching for academic resources over the internet and selecting the most appropriate and relevant ones.

Keywords: Critical Thinking, Information Accessibility, Search Engines.

1 Introduction

The estimated usage of smartphones users for the US in the 18-34 demographics is roughly 80%, and the main limitation to more common usage is cited as financial reasons [Rogowsky, 2013]. However, for those who have these devices information is now available anytime. With the modern idea of a Digital Citizen [Mossberger et al, 2011] who commonly uses technology in everyday life to engage in a variety of aspects of life, it is easy to see why current research promoting technology has a tendency to view computer usage as natural and does not impede the thinking process. But this new digital identity can make for a more solitary environment where the student may not question the information presented to them and diminish the skills associated with critical thinking.

How can we alleviate this problem whilst accounting for technologies growth in usage and acceptance as a major form of research? The need is simple, a more accessible, accountable research environment. Up until 2001 the .edu top-level domain was available to identify educational institutions. In 2001 however, it was closed to a limited number of US only post-secondary institutions.

There are standard academic digital resources such as EBSCO, SAGE, Taylor & Francis online, etc. but these are not available to all without subscription. However, one of the main free resources, Google Scholar, available to all, is limited by its own algorithms. It has been shown to be susceptible to spamming [Beel & Gripp, 2010] and occasionally it provides incorrect results [Jasco 2010].

Ideally an independent educational Top Level Domain that is available to academic institutions, libraries, publishers & researchers is needed. However, this would be an unattainable goal as the cost

to move already well-established institutions, including digital institutions like MOOCs and open universities would be astronomical.

What is achievable is creating an environmental link using the resources that are already available. An academic backlink system, providing a linking and rating system globally to allow for a controlled environment, maintained independently would be useful. The backlink system should be tailored to each of the various academic resources that could in turn feed into an independent academic search engine. Institutions could be responsible to ally the code to their own digital content whilst an independent administrator could assure all resource categories are appropriately adhered to. This search engine would not only allow for a stable academic research environment but also would enable simple apps to be used with standard search engines to identify content from this academic system minimizing cultural change in students search patterns.

2 Critical Thinking

Dewey referred to critical thinking as "active, persistent, and careful consideration of any belief or supposed form of knowledge in the light of the grounds that support it and the further conclusions to which it tends" [Dewey, 1933], a view that he expands on as a sense of uncertainty and suspense as necessary elements for critical thinking to take place. Dewey does not give a step-by-step process but rather proposes that it is an ideal of relationships forming to ensure learning takes place. Glaser in contrast defined critical thinking as a group of three elements [Glaser, 1941]

1. An attitude of being disposed to consider in a thoughtful way, the problems and the subjects that come within the range of one's experiences,
2. knowledge of the methods of logical inquiry and reasoning, and
3. some skills in applying those methods.

Glaser is notably building upon Dewey's original definition but has added much more focus to the proposed "methods of logical inquiry and reasoning" with "skill". Glaser has considered that these skills or "abilities" [Glaser, 1941, p. 6] are necessary and the desires to use these skills are fundamental, which is not accounted for in how we currently use technology.

3 Digital Resources for Critical Thinking

The World Wide Web brought a new challenge to critical thinking as standards across information providers do not exist. What the research presented in this paper considers to be an unrecognized threat is how information is decided to be relevant. Many students use the internet in conducting research. In fact the Internet is perceived by the learners as the only source in conducting research [Graham & Metaxas, 2003]. These view were substantiated by Sampath Kumar & Kumar who in their study found 91.93 percent used google for information retrieval [Sampath Kumar & Kumar, 2013]. According to Alexa.com (a widely trusted subsidiary company of Amazon.com that provides web traffic data analysis), the top ranked websites used for retrieving information as of March 13th 2014 are:

1. **google.com** - Enables users to search the world's information
2. **facebook.com -** A social utility that connects people, to keep up with friends
3. **youtube.com** - YouTube is a way to get your videos to the people
4. **yahoo.com** - A major internet portal offering search results
5. **baidu.com** - The leading Chinese language search engine
6. **wikipedia.org -** A free encyclopedia built collaboratively using wiki software.

Are these the primary sources for information? These search engines use sites that focus on SEO (Search Engine Optimization) techniques to assure their visibility. That is to say the site with the best advertising and the site structure (back links, reciprocal links, meta tags, etc.), not necessarily the best or most reliable information, will appear as a higher result. Graham and Metaxas's research [Graham & Metaxas, 2003]also showed that 48% of students in one case believed an incorrect piece of

information entirely sourced from the one website and they confidently cited it without feeling the need to find a secondary or corroborating source. Their research highlights the concern that students need to develop critical thinking skills earlier in order to be able to identify reliable sources and to truly reap the benefits of these online research methods.

The online resources are helping to shape this Digital Citizenship and learning identity, "todays citizens are expected to attend social processes anytime and anywhere" and on a higher level need to develop a Digital Identity [Simsek & Simsek, 2013]. Convenience is the key and mobile technologies have made this more abundant than ever. Therefore, there is a need for forward planning to provide a scenario for growth in materials and resources, yet control over visibility of relevant and reliable information.

Prior studies have investigated the area of search engine functionality. And although many can see that there is a change in the thinking process of students with these tools the approaches seem to be of a more personal approach. These studies suggest that an intensive training for academics needs to be conducted to acquire search strategy skills [Sampath Kumar & Kumar, 2013] or that guidance and support should be provided for use of information retrieval systems [Brophey & Bawden, 2005]. These seem to be the more common approach especially in recent studies. The optimization of search functionality for academic has been explored before. The Bielefeld Academic Search Engine was created to be a scholarly research tool [Pieper & Summann, 2006]. BASE in face still functions today, but is in itself limited to members of its own search pool and requires it be used as the primary search tool.

4 Proposed Research

The proposed research would investigate the features/specifications of the existing academic resources and will propose a methodology for ranking these resources based on how relevant they are for the learner. The goal of the research is to propose a solution that supports and improves critical thinking backlink code for existing academic resources and a tool to identify these materials and resources will be developed. A Learner will be able to use a searchable site consisting of the selected materials, and a browser plugin to identify materials from relevant sources.
For the purposes of initial research Google Chrome browser and google.com as the primary search for the plugin will be used. This is an aesthetic choice by the researcher as Google.com is the most used search online engine and Chrome is the standard browser for all android mobile devices.

This research study will firstly identify what are the clear classifications of academic resources and appropriate labels to apply to each one. e.g. peer-reviewed/not peer-reviewed, cited / not cited, reputation (based on authors name, authors institution, publication venue) A learner model that stores details about the learner such as interest, knowledge level, age, etc. will be designed. A context model will consider learner location, learner device type and network connectivity.

Then, a search algorithm will be proposed to weight relevance in academic resources search based on the labels assigned to the academic resources, on the learner model and on the context model. For example, if the learner is interested in searching for resources related to the 3D videos keywords while he is located in the College, using a desktop PC, he is actually looking for educational type information on 3D video (e.g. features of the 3D technology, research publications on this topic). As he is using a desktop, qood quality video clip type information can be suggested to him. If the student is using his 3D enabled HD TV set and the UPC Horizon device from home and searches over the Internet for the same keywords, he might be interested in entertainment type 3D video clips to be played on his TV. Since he uses a HD TV to watch the videos, HD, 3D format videos links can be suggested to him

Secondly a backlink system would need to be created to identify academic resources from participants in this system. These back links would feed into a site and stored in a database.

Finally the searchable interface of the site and the plugin app would be created. These applications will be used to test if the proposed ranking methodology and the search algorithm help the students by providing a better selection or identification of reliable materials. In turn, this will make students to question the sources of these materials and think more about what they are using. A tool to both support and improve critical thinking.

5 Conclusions

The results of this research have clear implications for the sourcing of information and providing an understanding of the relevance of materials used for students. A methodology for ranking and sourcing academic material is more necessary than ever. Internet resources are becoming more abundant and readily available, and the need for a tool that can identify an academic resource, that is more relevant and that utilizes these existing systems, is needed to better support the learners in searching for relevant information as well as to improve critical thinking skills. This will allow students to utilize the resources they are familiar with, with potentially greater efficiency and results.

6 References

[Beel & Gipp, 2010] Beel, J. & Gipp,B. (2010). On the Robustness of Google Scholar Against Spam. *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia*,10:297-298.

[Brophey & Bawden, 2005]. Brophy, J. & Bawden, D. (2005) Is Google enough? Comparison of an internet search engine with academic library resources, *Aslib Proceedings, Vol. 57*, 6:498 – 512

[Dewey, 1993] Dewey, J. (1933). *How we think*. Lexington, MA: D.C. Heath & Co.

[Glaser, 1941] Glaser, E. M. (1941). *An Experiment in the Development of Critical Thinking*. New York, Bureau of Publications, Teachers College, Columbia University.

[Graham & Metaxes, 2003] Graham,L., & Metaxas, P. (2003) Of course it's true; I saw it on the internet! Critical thinking in the internet era, *Communications of the ACM, Vol 46*, 5:70-75.

[Jasco, 2010] Jacso, P. (2010) Metadata mega mess in Google Scholar, *Online Information Review* 34:175–191.

[Mossberger et al, 2011] Mossberger, K., Tolbert, C., McNeal, R. (2011) *Digital Citizenship. The Internet, Society and Participation*. The MIT Press Cambridge, Massachusetts, London, England.

[Pieper & Summann, 2006] Pieper, D. & Summann, F. (2006) Bielefeld Academic Search Engine (BASE): An end-user oriented institutional repository search service, *Library Hi Tech, Vol. 24*, 4:614 - 619

[Rogowsky, 2013] Rogowsky, M. (2013), More Than Half Of Us Have Smartphones, Giving Apple And Google Much To Smile About, *Forbes*, 6 June 2013. Web 15 March 2013, http://www.forbes.com/sites/markrogowsky/2013/06/06/more-than-half-of-us-have-smartphones-giving-apple-and-google-much-to-smile-about/

[Sampath Kumar & Kumar, 2013]Sampath Kumar, B.T. & Kumar, G.T. (2013) Search engines and their search strategies: the effective use by Indian academics, *Program: electronic library and information systems Vol. 47*, 4:437 – 449

[Simsek & Simsek, 2013] Simsek, E., & Simsek, A., (2013) New Literacies for Digital Citizenship. *Contemporary Educational Technology Vol 4*, 3:126-137

Session 4

Research in Progress Pitch & Doctoral Symposium

Thesis 1

Title: QoE derived Olfaction-enhanced Multimedia Synchronization

Author: Niall Murray

Supervisors: Dr. Brian Lee, Dr. Yuansong Qiao and Dr. Gabriel-Miro Muntean

Abstract:

Recently, significant efforts have been made by researchers and industry to propose solutions to increase multimedia viewers' QoE. One approach is to use media components beyond the traditional audio and video, but to understand and model users' perception of mulsemedia experiences is a multi-dimensional and challenging problem. In this context, this PhD thesis further advances the understanding of users' perception of olfaction-enhanced multimedia by considering the impact of various network transmission effects. The contributions are:

- A subjective study was carried out to investigate user's detection and perception of inter-media skew, and the effect skew has on users' QoE. The results report that users are tolerable to certain levels of skew before deterioration in QoE occurs. In addition, differences in detection and perception of skew exist considering user context (age, gender and culture i.e. user profile).

- The users' perception of typical technical network characteristics such as delay and jitter between two olfactory media components was captured during a second subjective study. The effect of varying the temporal distance between the release of two olfactory streams was analysed considering: user ability to detect skew, user perception of skew, the effect on user QoE, the minimum duration required between their release to ensure enhanced QoE and user perception of "mixing" of different scent types.

- A model to estimate user QoE of olfaction-enhanced multimedia was researched and developed. It considers the influence of inter-media skew, users' context (profile) and scent type as input criteria and validates the model against the results of subjective studies.

- A comprehensive review of the state of art on the use of olfaction across a number of industries was performed. We also provide a unique set of guidelines, based on the expertise gained, for the execution of subjective testing involving olfaction as a media component.

Thesis 2

Title: A Cross-layer Content-aware Energy-efficient Rich Media Application Delivery Scheme in Heterogeneous Wireless Networks

Author: Shengyang Chen

Supervisor: Dr. Gabriel-Miro Muntean

Abstract:

As one of the most state-of-the-art and popular communication technologies, the development of wireless communication technology has been focusing on fulfilling the demand in various parts of human life over the last decades. In many real-life cases, this demand directs to most types of commonly-used rich-media applications which often require considerably high quality levels on the end devices of wireless network users. Different types of rich-media applications have diverse traffic patterns individually and applications with certain patterns might be suitable to be transferred through one type of wireless network but not others. For this reason, deliveries of such applications are usually accomplished using heterogeneous wireless networks that use more than one type of wireless network structures simultaneously. On the other hand, content deliveries with assuring quality involves increased energy consumption on both ends of the communication and thus highly challenges the limited power resource of the devices in the wireless networks, especially battery-powered end devices of users. As a result, many research and industrial efforts have been invested aiming at high-quality and energy-efficient rich-media content deliveries in the past years.

Our research presented in the thesis focuses on energy-aware content delivery scheme in heterogeneous wireless networks, including wireless LANs and cellular networks. It mainly contributes in the following aspects:

- E-Mesh - an energy-efficient cross-layer solution for high-quality multimedia content delivery over wireless mesh networks which provides an adaptive device sleep-period management scheme at the Medium Access Control (MAC) layer together with an energy-load-distance-aware routing algorithm at the transport layer. The novelty of E-Mesh is the cooperation between the MAC-layer scheme which adaptively controls the sleep/awake pattern for mesh devices and the transport-layer routing algorithm which takes energy consumption, change of device position and load on devices into consideration at the same time. The benefit of E-Mesh is that the quality stability in a variable mesh network environment is maintained by the routing algorithm so that the possible fluctuation of traffic that might seriously affect the multimedia delivery quality is avoided. Meanwhile the battery standby time for mesh devices is extended by introducing the sleep-period management scheme which allows idle devices to be turned off to save energy.
- eMTCP - an energy-aware MPTCP-based content delivery scheme in heterogeneous wireless networks which provides a traffic offloading scheme for multiple wireless network interface usage on mobile devices. The novelty of eMTCP is that it takes into consideration of energy efficiency when using different wireless network communication technologies such as LTE and Wi-Fi at the same time, based on the MPTCP protocol which significantly improves content delivery quality. The benefit of the eMTCP is that it balances the support for increased content delivery quality from MPTCP with energy consumption awareness, without

any additional modifications or deployments on the wireless device structure. Meanwhile the energy consumption levels for different types of wireless network interface are investigated and an optimal traffic offload amount for achieving best quality is evaluated.

- eMTCP-BT – an extension of eMTCP that determines the optimal traffic offload percentage for diverse types of application with different traffic patterns. It uses mathematical models such as decision tree or Markov decision process to analyze traffic pattern in order to make traffic classification and to adapt the corresponding optimal offload percentage based on the class of traffic.
- A combined solution for the previous schemes is proposed which takes benefits from both schemes for further increasing the service quality for wireless network users.

Session 5

ICT in Education

Comparing Game Based Learning, using a student created game, to Traditional Classroom Methods

Jeremy Rigney[1], Niall Murray[1]
[1] Athlone Institute of Technology, Ireland
jeremy.rigney@gmail.com, nmurray@research.ait.ie

Abstract

This paper presents an initial study that compared the effectiveness of Game Based Learning (using a student designed game) with traditional classroom methods. The task was to teach students a basic math concept. Previous works focused on the use of subject specific mass marketed games. Here, the game developed was based on free software. The results indicate that the student created game is as effective as Traditional Classroom methods, but more efficient. The results also demonstrate that Game Based Learning could be equally effective for both male and female students. We conclude that student created games, can be employed in addition to current teaching methods, providing greater student engagement.

Keywords: Game Based Learning, ICT, education, Serious Gaming

1. Introduction

Game Based Learning (GBL) is the use of games as teaching tools, in place of, or as an enhancement to, the traditional classroom (TC) settings. In many learning environments games are already used for entertainment value. However, GBL is seen as an approach with significant potential to translate entertainment into education [1], an approach often presented as edutainment. In recent times, games have been introduced as learning tools in schools on trial basis. Their full potential is not yet understood. There are varying opinions on the use of GBL. Some researchers suggest it promotes aggressive behavior [2], whilst others [3] propose that GBL will become more widely used in education.

Felicia in [1] presents the current classroom environment as being teacher, rather than student, centered. From the student perspective, they are only passively engaged, whilst teachers are the focal point in the learning experience. Passive learning involves the passive absorption of information, listening and note taking. In contrast, active learning involves 'learning by doing' where 'hands on' work is used to develop skills [4]. GBL falls into the 'active' learning category.

This paper reports a student created 'Drill and Practice' game. The game poses a number of questions to the player. As the player develops their skills, the complexity and challenge increases. The effectiveness of Drill and Practice games in comparison to other educational game styles is an open research question. While some believe that they do not support learning, others feel it is as effective as larger more complex games. While there is a large amount of empirical data regarding the use of Simulation games, 3D immersive games, mass marketed games and Massively Multiplayer Online Role Playing Games (MMORPGs), little works have reported on the evaluation of educational games created by students. In this context, this study investigates if a student designed educational computer game is as effective compared with traditional classroom methods, to teach a basic maths concept.

2. Related Work

GBL fosters several benefits such as (a) supporting multi-sensory, active, experimental problem-based learning, (b) activation of prior knowledge given that players must use previously learned information in order to advance (c) immediate feedback enabling players to test hypotheses and learn from their actions [7], and (d) opportunities for self-assessment through the mechanisms of scoring and reaching different levels [3], [8], [9], [10]. It has been proposed that both intrinsic and extrinsic motivators are key advantages of educational video games. Game environments are intrinsically motivating in that participation in the game is its own reward [11], [12], [13]. Students also benefit from the immediate constructive feedback provided in GBL via positive externally motivating mechanisms, such as scoring and levels [14], [7], [8], [12]. GBL can promote the fun and engaging properties of entertainment games while allowing the participants to learn.

'Flow' is a concept developed by Csikszentmihalyi in [15] to describe the experience of total immersion in an activity to the exclusion of the immediate environment. In terms of game play, the benefits of achieving flow include an increase in focus and concentration together with an increase student engagement, all of which are strongly associated with student achievement [7]. Game components such as (a) clear goals (b) direct and immediate feedback (c) balance between ability level and challenge, and (d) sense of control, are constituent elements of the flow experience.

Egenfeld in [16] highlights a number of issues with GBL within the educational setting such as short lessons, physical space, variations in game competence among students, installation, costs, and teacher preparation time. Also it has been claimed that games are a 'male dominated domain' with males more frequent and intensive game players than females [17].

The role of the teacher is crucial with the integration of information technology in schools, but they have also been considered one of the main barriers [17]. The use of GBL has focused of Mass Marketed Subject Specific (MMSS) games, Commercial Off the Shelf Games (COTS) [6], and MMORPGs. But, with such approaches teachers need to adapt and develop lesson plans based on the game. Little scope exists to adapt the game to their own objectives. The use of GBL also requires teachers to have access to computers with 'gaming specification', technical support, familiarity with GBL content, adequate preparation time and the cost of licenses for access to educational games [18]. All of these can be a deterrent to the implementation of GBL.

In contrast, the 'Scratch' software used in this experiment is a free piece of software, supported with free online lessons, thus making it easier for teachers and students to create simple yet relevant classroom material. In this context therefore, the use of a student designed game is an example of how this software might be successfully applied as an enhancement to traditional classroom teaching methods. Additionally the simplicity of the software provides the foundation to explore 'peer-to-peer' learning with students actively involved in designed game based learning projects.

This paper extends state of the art by employing free software and online tutorials that would allow students and teachers to create their own games. Where there is now a large body of research investigating the use of mass marketed games in education, a gap remains regarding the use of student created games for learning [19]. This study compares the effectiveness of Game Based Learning, using a student created game, to traditional classroom methods to teach a basic maths concept.

3. Experimental Design

3.1 Participants

50 participants (24 male and 26 female) were recruited with age ranges from 10yrs-12yrs. These were split into the control and test groups. Teacher opinion suggested that all participants were inexperienced in Prime and Composite numbers. Thus this was the topic chosen to teach through the game. Consent for photography and video was obtained prior to the experiment from the teacher. Permission was also obtained from parents prior to any testing.

3.2 Materials

Scratch version 1.0 software was used to build the game. Twenty five Samsung and Toshiba Netbooks supplied by the school and used by the experimental group to play the game. The specifications of the Samsung Netbooks were: Samsung Netbook Model np-n145, RAM 1 GB, Processor; Intel Atom CPU n455 @1.67Ghz and Toshiba Netbook Model nb250-108, RAM 1 GB, Processor; Intel Atom CPU n455 @1.67Ghz, both with Windows 7 Starter 32bit. Scratch was installed on all computers. One Toshiba Laptop was used to connect to a projector to show a distraction video. The distraction video was used to ensure that the cycle of learning was broken before the post-test was administered. This meant that the results of the post-test were the result of learning and not short term retention.

Two classrooms were needed to ensure both groups did not mix. Participants were required to complete questionnaires and pre and post-tests. The Questionnaire was administered to the Experimental group to understand whether the participants in that condition enjoyed the game and their opinion on whether GBL should be implemented in schools.

3.3 Game Design

The design of the game followed the 'General Game Design Elements' as described by Shabanah in [5]. The participant is asked a question, for example "Which number is Prime". At the same time, four numbers appear on small glass windows. The participant types in which number they think is Prime and a chicken is launched from a catapult, breaking the window their number is on. If their answer is correct the score is increased by ten. If incorrect it decreases by ten. A system of rewards and goals to motivate players was employed to stimulate desired learning outcomes [12]. The game began with seven slides to introduce the topic. This was followed a practice section and consisted of twenty questions, including bonus question. If a score of 100 is reached at any point throughout the game, the background of the game switches from day to evening. This feature was implemented to give the player a reward to encourage them to continue and to acknowledge their score. The game had a fun 'scribble page' at the conclusion as a reward for reaching the end of the game and as an overall goal to try and reach. The game consisted of a scoring mechanism to provide real time feedback in response to correct or incorrect answers. Cassel and Jenkins [17] state that generally game designs appeal mainly to boys. Therefore the game used in this experiment was designed with a gender neutral 'fun' theme in mind (e.g. no shooting and the use of a chicken rather than a gender specific person). The game was designed to be colourful, eye catching and inviting.

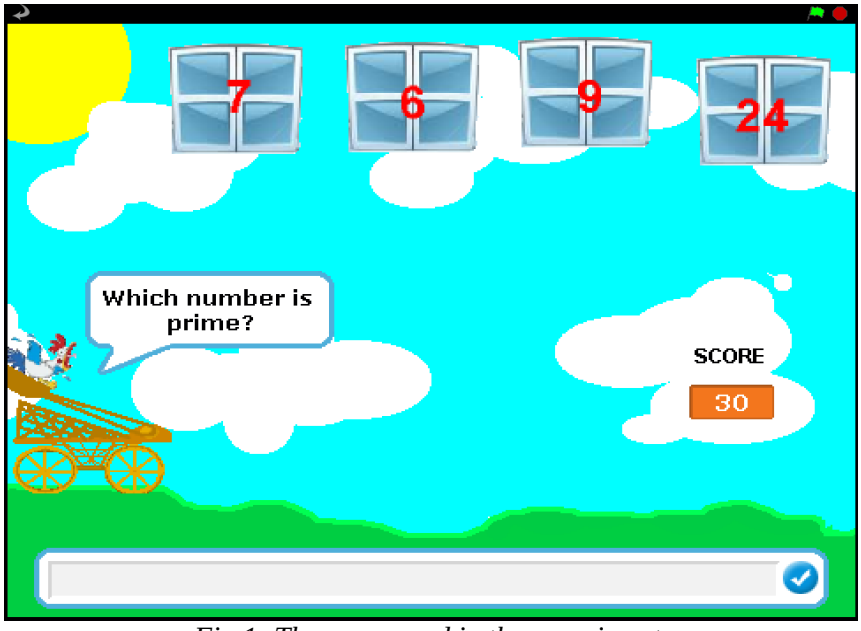


Fig.1: The game used in the experiment.

3.4 Procedure

The learning topic 'Prime and Composite' numbers was chosen via collaboration with a second level maths teacher. None of the participants had received explicit instruction in this topic prior to the experiment. A number of steps were taken to minimise confounding variables; the experiment was conducted after lunch at 1:45 to control hunger, fatigue and boredom. Additionally the two conditions, GBL and TC, ran simultaneously in adjacent rooms to ensure that environmental factors such as light and heat were the same.

A between participant design was used in this experiment. Participants were randomly assigned to one of two groups; the TC group (control) or the GBL group (experimental). The groups were then brought to their allotted classrooms and given a written pre-test on the topic of prime and composite numbers. The time allowed for completion was ten minutes. This was used as a base against which to assess learning. The control group (TC) received instructions on prime and composite numbers from a teacher via traditional classroom methods. Twenty five minutes was allocated for the teaching of the topic in this condition. The second group (GBL condition) received instructions on prime and composite numbers via the game. Participants were provided with information on how to operate the game by the experimenter and were then allocated 25 minutes to complete the game. Given the wide range of ability levels and experience among pupils in a traditional classroom, predicting the required game completion time was difficult. Therefore an allotted time of 25 mins was allowed based on advance conversation with teachers familiar with their students' experience and ability. A teacher was present to supervise the room but took no role in instructing students. At the end of the 25 minutes both participant conditions were brought to one classroom where an interactive white board was set up They were then shown a short video for five minutes to distract and break the cycle of learning and to ensure that the results from the post-test were not due to short term retention of the topic.

After the video was shown, all participants were brought back to their test rooms where a post test was administered. Ten minutes was again allotted to complete the exam. Students in the GBL condition were also asked to complete a short questionnaire to gather general feedback about the game. When the tests were collected participants were gathered in one room, debriefed, given the opportunity to withdraw from the final results and were thanked for their time. Those in the TC condition were advised that an opportunity would be provided to play the game the following day.
The hypothesis under consideration in the present study is that Game Based Learning will be as effective as the Traditional Classroom methods to teach a basic maths concept. The null hypothesis will be that Game Based Learning will have no effect on learning outcomes.

The independent variable was the teaching method. The dependent variable was the difference between pre and post test results, i.e. the learning gains, by the participants. The mean difference in pre and post test results was used as the measure of learning. The results are presented in section 4.

4. Results and Discussion

While test results show similar learning gains in both GBL and TC conditions, the amount of time for task completion differed significantly between the conditions. The time allocated for completion was 25 minutes in both game (GBL) and class (TC) environments. While participants in the TC condition used the full twenty five minute time limit, participants assigned to the GBL condition, and who were allowed to continue at their own pace, completed the game in an average of ten minutes. This significant time difference suggests that the use of GBL may result in some concepts being taught faster and more efficiently then TC environments. This is a key result of this experiment. The initial hypothesis was that GBL would be as effective as the TC Methods, and this was the outcome of the experiment in terms of the results of the pre and post-tests.

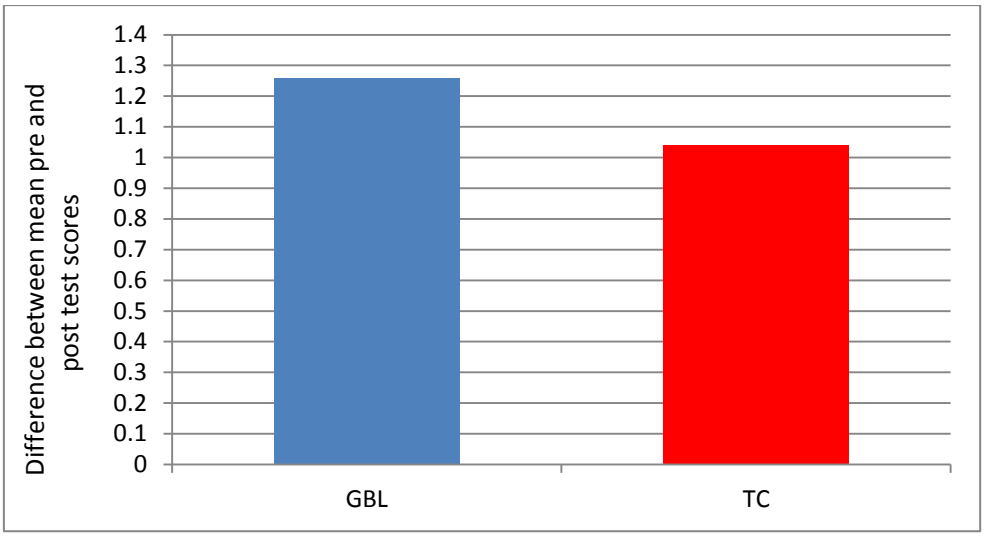


Fig.1: The mean difference in the increase between pre and post test scores in both GBL and TC conditions (outliers removed)

	N	Mean	Std. Deviation	Std. Error Mean
GBL	23	1.26	2.73	0.56987
TC	23	1.04	2.77	0.55758

Table 1: The mean difference, standard deviation and standard error (outliers removed)

There was no significant difference in the scores for GBL ($\bar{x}$ =1.26, SD=2.73) and TC ($\bar{x}$ =1.04, SD=2.77) conditions; t (48), p= (0.765) with outliers removed. These results indicate that Game Based Learning, using a student designed game, to traditional classroom methods is as effective. In addition, the difference in completion time between both conditions is an unexpected, yet significant, outcome of the experiment. It shows the potential for GBL to teach topics more efficiently than the TC methods.

Analysis of the data considering gender presents interesting findings in the context of previously accepted trend that games are more suitable for and are male centered. Gender comparison shows that females made marginally greater gains in learning than males in the GBL condition (females mean increase: 2.78, males mean increase: 2.5.) This suggest that GBL, as suggested by [17][16], can be as beneficial to both gender groups. Figures 2 and 3 present the mean increase with respect to learning for the GBL and TC.

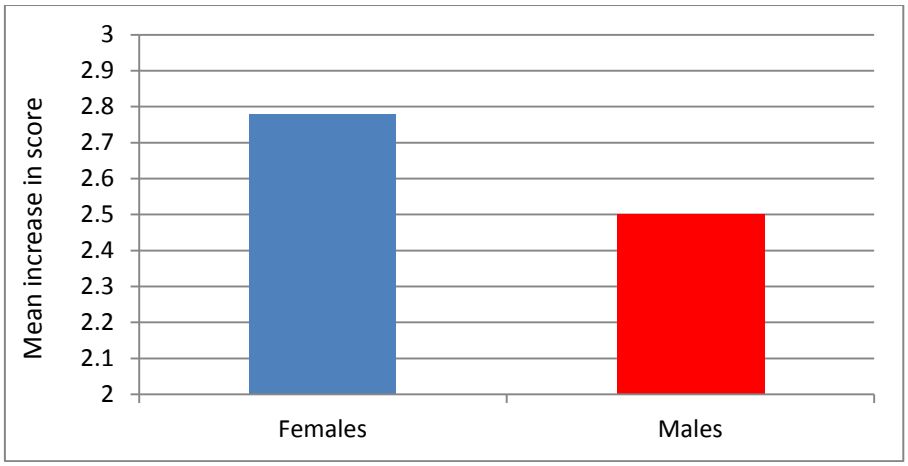


Fig 2. The mean increase in the scores between males and females in the GBL condition

A number of observations emerged from the experiment. Firstly, this research supports Felicia[1] in terms of student centered classrooms. The participants demonstrated active involvement in their own learning by engaging with the game, making comparisons between their own progress and that of other students with a competitive environment developing as students attempted to maximize their scores. The concepts of motivation and flow were also shown by the participants with the motivation to improve on their previous score and losing track of their environment and becoming entirely focused on the task at hand. In contrast to teacher opinion however, it was found that the participants appeared to have a basic grasp of the topic prior to the experiment, judging by their high pre-test scores. Future research may use a pre-test screening process to eliminate any participants who have a basic understanding of the topic.

5. Conclusions and Recommendations

This experiment compared the usefulness of GBL, using a student designed game, to traditional classroom methods to teach a basic maths concept (prime and composite numbers). The results of this experiment indicate that in terms of learning, a student designed game for Game Based Learning was as effective as the Traditional Classroom methods. Additionally the results indicate that GBL is equally effective for both boys and girls. These findings strongly suggest that incorporating GBL into a traditional classroom environment would be a positive enhancement to the learning experience. The main finding of this experiment was that there was a difference in overall completion times with those in the GBL condition taking significantly less time than in the TC environment. This shows that topics could be more efficiently taught through GBL.

GBL can be more efficient than TC for teachers because it can allow them to quickly gauge students' ability level, based on their scores and how fast they finish the games.

GBL is more efficient for students because they can work at their own pace; if they are of a higher ability they can complete the game quickly and move on to a new topic. If they are of a lower ability they can re peat the game multiple times until they feel that they have grasped the topic.

While previous research has focused on the use of mass marketed subject specific games, this experiment used the Scratch programming software which is a free, well supported package that allows both students and teachers to easily create games. Teachers in particular can adapt Scratch to suit their own teaching style rather than have to adapt their teaching style to a mass marketed game. Specifically, the use of software such as Scratch has the potential to overcome the cost, training and support issues associated with mass marketed games that can act as barriers to the implementation of GBL in a classroom setting.

A number of areas for future research were identified:

- More detailed experiments on efficiency of game based learning compared to traditional classroom methods.
- Are different subjects more effectively taught through GBL? (e.g. Arts or Science)
- Which age group or which learners GBL might be most effective for.
- Can students learn by creating the content without any formal instruction in the subject?

Acknowledgements:
The authors would to acknowledge Dr. Brian Lee for the use of facilities in the Software Research Institute in Athlone Institute of Technology. In addition they would like to thank the Principal and teachers of Durrow National School for the use of their school facilities and for their time, and their students in fifth and sixth class for participating in the Experiment.

References

[1] Felicia, P. (2011) "What evidence is there that digital games can be better than traditional methods to motivate and teach students?" *EUN Partnership AISBL*

[2] Provenzo, E.F. (1991). "Video kids: Making sense of Nintendo." Cambridge, MA: Harvard. *http://doczine.com/bigdata/1/1369989846_0d2af3a0f9/39-squire-ijis.pdf*

[3] Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K. (2011) “The 2011 Horizon Report” *The New Media Consortium*, ISBN 978-0-9828290-5-9
[4] Felder, R. (1993) "Reaching the Second Tier: Learning and Teaching Styles in College Science Education." *J. College Science Teaching*, 23(5), pp. 286-290
http://www2.ncsu.edu/unity/lockers/users/f/felder/public/Papers/Secondtier.html
[5] Shabanah, S.S, Chen, J., Wechsler, H., Carr, D., Wegman, E. (2010) “Designing Computer Games to Teach Algorithms” Proceedings of the Seventh International Conference on Information Technology, pp. 1119-1126
[6] Van Eck, R. (2006) “Digital Game-Based Learning: It's Not Just the Digital Natives Who Are Restless” *EDUCAUSE Review*, 41(2), pp. 16-30
[7] Larsen McClarty, K., Orr, A., Frey, P.M., Dolan, R.P., Vassileva, V., McVay, A. (June 2011) “A Literature Review of Gaming in Education”
[8] Tang, S., Hanneghan, M., El-Rhalibi, A. (2009), “Introduction to Game-Based Learning,” in Games-Based Learning Advancements for Multisensory Human Computer Interfaces: Techniques and Effective Practices”, T. M. Cronnolly, M. H. Stansfield, and L. Boyle, Eds., *Information Science Reference*, ISBN13:9781605663609, pp. 1-17
[9] Papastergiou, M. (2009) “Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation.” *Computers & Education* 52, (1), pp. 1–12
[10] Oblinger O.G. (May, 2004) “The Next Generation of Educational Engagement” *Journal of Interactive Media in Education, Special Issue on the Educational Semantic Web* ISSN: 1365-893X
[11] Prensky, M. (2001) Digital Game-Based Learning McGraw-Hill, 2001
[12] Dondlinger, M.J. (2007) “Educational Video Game Design: A Review of the Literature” *Journal of Applied Educational Technology* vol.4, no.1
[13] Rieber, L. P. (1996) “Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games” *Educational Technology Research & Development*, vol.44, no.2, pp.43-58
[14] Liao, Y. (2011) “Game-based learning verses traditional instruction on student affective outcomes in Taiwan: A meta-analysis.” *Journal of Information Technology and Applications*, vol.5, no.1, pp. 28-36
[15] Csikszentmihalyi, M. (1990) “Flow: The Psychology of Optimal Experience”
[16] Egenfeld-Nielsen, S. (2006). ‘Overview of research on the educational use of video games’, Nordic Journal of Digital Literacy, http://www.idunn.no/ts/dk?languageId=2
[17] Cassel, J., & Jenkins, H. (Eds.). (1998). From Barbie to mortal kombat: Gender and computer games. Cambridge, MA: MIT Press. pp. 298-325
[18] de Freitas, S., Jarvis, S. (2007) “Serious games--engaging training solutions: A research and development project for supporting training needs.” *British Journal of Educational Technology*, vol.38, no.3, 2007, pp. 523-525
[19] Prensky, M. (2008) “Students as designers and creators of educational computer games: Who else?” *British Journal of Educational Technology*, vol.39 no.6

Practice Testing - Enhancing Student Learning with E-Assessment

James Eustace, Pramod Pathak, Cristina Hava Muntean
National College of Ireland, IFSC, Mayor St, Dublin 1
james.eustace1@student.ncirl.ie, Pramod.Pathak@ncirl.ie,
Cristina.Muntean@ncirl.ie

Abstract

Many students struggle to regulate their learning, often using ineffective techniques. The utility of practice testing and distributed practice is supported by a large body of research from the cognitive and educational psychology community. Computer-based testing (e-assessment) has had a long history with many opponents yet continues to be used in a wide range of educational contexts and professional certification. Given the robustness of the testing effect phenomenon, e-assessment has the potential to transform the student self-regulation landscape. The materials and test items employed in practice testing have not explored fully the boundaries of the testing effect in areas involving transfer of learning and higher order thinking. The boundaries between practice testing and formative assessment are subtle but distinct. The aim of this paper is to provide a survey/state of the art on practice testing by analysing the methodologies employed by the practice testing community and explore how existing e-assessment technologies can be employed to extend the research on practice testing. The challenge for technologists and educators alike is to leverage the potential of e-assessment for higher order learning and not simply replicate traditional paper-based tests in an electronic format.

Keywords: E-learning, Testing effect, Computer-based testing, Question and Test Interoperability

1 Introduction

Research has shown that many students struggle to regulate their own learning, often using ineffective techniques such as rereading or highlighting. Techniques such as practice testing and distributed practice have been shown to be highly effective across a range of conditions and criterion tasks and are supported by a large body of research [13]. Practice testing, also referred to as retrieval practice, benefits learning where material is better remembered over time if the learning process includes tests on that material. While the techniques used in formative assessment can be used to enhance learning they are not as a result of the direct effects of testing on learning but represent the mediated effects of testing on learning [30]. The testing effect phenomenon occurs as a direct result of practice testing which includes any form of practice testing such as low stakes, no stakes or self-testing which learners can engage with on their own. A testing effect is evident if performance on the final test is greater in the test-study condition than in the study-only condition. Different theories have evolved to explain the testing effect. The overlearning or increased exposure hypothesis [34] argues that the testing effect is due to overlearning of the items through repeated exposure and successful recalling. However the testing effect cannot be fully explained by the amount of exposure alone. The elaborative retrieval hypothesis [23] advances the idea that the process of retrieval modifies memory and increases the probability of future successful retrieval. An alternative hypothesis looks at the testing effect from a transfer-appropriate processing viewpoint where memory performance is dependent on the overlap between the encoding process and the retrieval process [24]. Within this framework it is the student engagement with similar operations during testing that results in enhanced performance compared with items not tested or only restudied.

Different methods have been employed to investigate practice testing. In general, the experimental conditions find students being presented with material to be learned during an initial learning phase.

The students are typically assigned to groups where one group completes a practice test or a sequence of practice tests (TTT) while the other group studies the material again (SSS) as illustrated in Figure 1. Learner performance on an immediate test and subsequent final delayed test is consistently greater following a practice test compared with having no practice test. The research findings support the claim that practice testing is more effective than restudying alone and that by combining practice tests with restudy over time is even more effective [6]. The research to investigate the testing effect is wide and varied yet is limited in its implementation in educational settings [27].

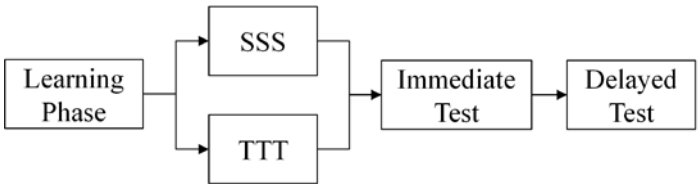


Figure 1: Typical practice testing procedure

This paper analyses the practice testing methodology and proposes an assessment design approach to extend the practice testing research by using e-assessment. Section 2 provides an analysis of the practice testing methodology and positions practice testing within educational assessment while Section 3 analyses practice testing from a design approach within an e-assessment development framework and materials used. Section 4 concludes the paper with recommendations for future work.

2 Practice Testing and Educational Assessment

The literature on educational assessment can be divided by purpose into four overlapping areas; diagnostic, formative, summative and self-assessment. Diagnostic assessment is generally used to determine what the student already knows, to identify gaps or difficulties that may exist and is typically employed pre-instruction. An assessment is formative if the information provided is used to help shape or adapt approaches to student learning. Summative assessment is typically performed at the end of a period of learning and is used to determine the student's level of competence and to assign grades based on achievement of curricular objectives. These three roles for assessment are typically instructor driven and the information generated is used by instructors for different purposes as described. Self-assessment is a technique used by students to enhance their own learning, to confirm their understanding and grasp of a particular knowledge domain. For example, practice testing could involve students using flashcards to recall information or completing practice questions at the end of each chapter in school textbooks. Practice testing fits within the self-assessment role of educational assessment but can use design characteristics or elements of diagnostic, formative and summative assessment. Practice testing facilitates two key requirements for students, the active retrieval and use of relevant domain knowledge and being able to make judgments as to whether the performance standard has been achieved on the feedback received. When considering practice testing as an effective technique a brief review of the research methodology follows in the next section.

2.1 Practice Testing Research Methodology

An analysis of the methodologies employed regarding practice testing is presented next and highlights emerging issues as well as discussing advantages and disadvantages. A review of the research performed in this area shows the robust nature of the testing effect phenomenon across a wide cohort of students. While the participants for many of the studies involving practice testing were undergraduate students [26] experiments on testing effects have been conducted with participants selected from a wide range of age groups and backgrounds both academic and community based as cited in [13]. The number of participants in the practice testing literature vary with studies ranging for as few as twenty four [13] to studies with over three hundred participants [27].

The research design which follows describes the treatments or manipulations and the measurement procedures generally conducted. The most common experimental design approaches adopted in the

practice testing literature are the independent or between-participants factorial design [20] where participants are randomly and independently assigned to each level of the independent variables; and the dependent or within-participants factorial design [23] where the same participants are exposed to all treatment conditions. Mixed design approaches are also evident in the literature [27] where a variable or variables is/are assessed between-participants, and another variable or variables is/are assessed within-participants. The mixed design approach can help with carry over effects however statistical analysis tends to be complex. A review of the supplementary material provided in [27] indicated that many experimental designs had retention periods of a week or less and in many cases lasted only a matter of minutes. Designs which extended beyond a week employing educational relevant materials were limited. The majority of these studies were conducted in laboratory type settings using controlled experiments [26]. While this design approach allowed better control over the independent variables and permitted more precise measurement of the dependent variables, the outcomes may not be as applicable in real world situations.

A wide range of materials have been employed in the practice testing field. Many of the experiments involved materials such as text [5], facts [31]and more complex materials such video [10]. However while these materials have relevance in an educational context, in many cases did not reflect actual classroom curricular content. This has led to an increase in focus of studies using curricular content in educational settings and course material using actual tests [29]. The research provides limited examples of materials intended to assess higher order learning. In many studies the final test material is identical to the practice test [23], while a number of studies have used modified or rephrased versions of the same test [22], evidence of the testing effect is limited when the final test requires transfer of learning [32]. In a study involving problem solving the testing effect was not evident using worked examples. The materials were paper-based and required the participants to troubleshoot electrical circuit faults consisting of isomorphic test items which necessitated the application of solution procedures requiring inferences to be made and perhaps indicate a boundary for the testing effect [15].

The format of tests administered during practice tests typically employed a wide range of item types and mainly using paper and pencil tests. The most frequently employed format is cued recall where the participant recalls items which were previously presented during initial training with the number of items recalled providing a measure of learning. The use of multiple-choice items was found to have both positive and negative effects on learning performance however the provision of feedback reduced the negative aspects and enhanced the positive aspects [6]. A number of studies used online testing to enhance performance on assigned textbook reading [19] and compared in class quiz with online quiz [11] while other quizzes were administered using clicker response system [23].

The procedures for practice tests are typically broken down into sessions or phases. In the simplest form participants undergo a practice test or a number of practice tests compared with other study methods before the administration of a final or delayed retention test. Participants typically are randomly assigned to groups using a factorial design and provided with instructions to complete the assigned tasks. A typical sequence consists of a timed study phase, a timed testing phase, a timed filler task and a final cued recall test [31]. In many of the experimental procedures much of the research has been focused on establishing the efficacy of practice tests over restudy and no practice tests over the same time frame. The research needs to extend further to determine the boundaries of the testing effect in facilitating higher order learning in areas such as problem solving.

2.2 Practice Testing and Higher Order Learning

When assessing learning within any complex domain no single test can measure the full range of outcomes. A combination of different assessment techniques are required that are valid, reliable, fair and practical to administer where online testing is one such technique. When developing learning objectives and assessing cognitive outcomes a taxonomy such as Bloom's taxonomy [4] is widely used. A revision of this taxonomy [1] differentiates between "knowing what," the content of thinking, and "knowing how," the procedures used in solving problems. This taxonomy is organised into six

different categories from the most basic, remembering to more complex levels of thinking. From a practice testing perspective declarative knowledge or remembering is fundamental to solving more complex problems as within the taxonomy each level is a prerequisite for the level above. Testing recall of declarative knowledge has been the focus of much of the practice testing research. There is general agreement that when assessing declarative knowledge and when assessing cognitive skills or abilities where there is a right answer, selected response and constructed response item formats are comparable [16].

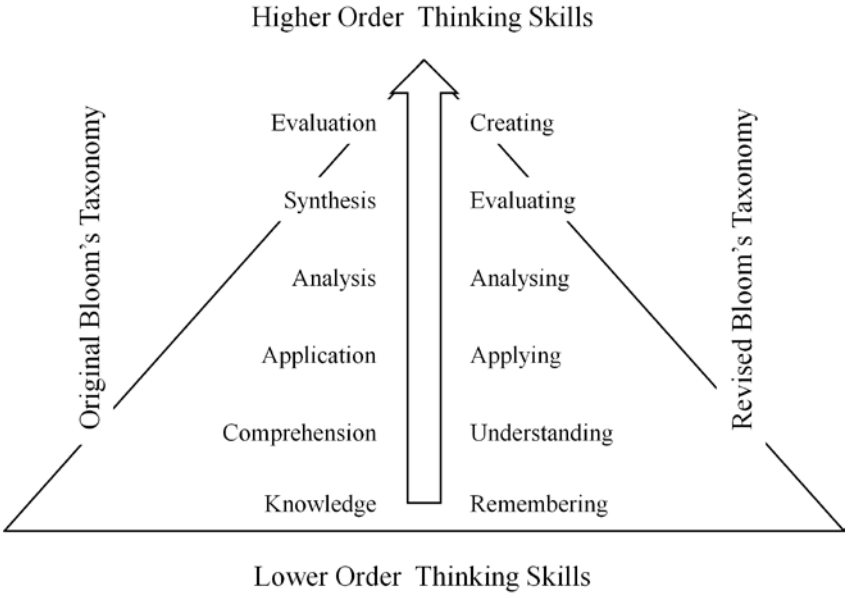


Figure 2: Bloom's/Revised Bloom's Taxonomy - Lower to Higher Order Thinking Skills

One of the primary goals of learning is the application of knowledge to new situations commonly known as transfer. Transfer research can be viewed broadly under three contexts, temporal, test format and knowledge domains [7]. The temporal context of transfer relates to how well information is remembered over time and is supported by a large body of practice testing research. Test format has an impact on the testing effect where test items that are more demanding, have a greater learning benefit. Findings show that when students are tested with higher level questions they acquire a deeper understanding for the course material [18]. For example the modular and sub modular outcomes within aviation technician training are defined in terms of these levels and are assessed in a number of ways including multiple-choice test items. The general definitions for each level is described in Table 1, for a more detailed definition refer to [14]. A conceptual mapping to the Taxonomy above would see Level 1 as a lower order thinking skill which increasing thinking skills through Levels 2 and 3.

Table 1: Knowledge Levels for Aircraft Maintenance Technicians

Level 1	*Level 2*	*Level 3*
A familiarisation with the principal elements of the subject.	A general knowledge of the theoretical and practical aspects of the subject and an ability to apply that knowledge.	A detailed knowledge of the theoretical and practical aspects of the subject and a capacity to combine and apply the separate elements of knowledge in a logical and comprehensive manner.

Problem solving is considered to a higher order thinking skill. For a person to solve a problem they must be aware of their initial state and their final or goal state, they must then identify and go through an operation or sequence of operations to arrive at their goal state which is referred to as the information-processing approach where the problem solver creates a mental representation or model of the problem in their head [33]. Online test items can be enhanced with multimedia elements within the question item to test problem solving skills [25] not possible in traditional paper and pencil tests using the same item formats. Problem solving can be divided into three core dimensions; task or problem

statement which provide the conditions or directions needed; technologies, the mechanism through which the problem solving is conducted and the cognitive dimensions, the cognitive structures and processes required to solve the problem [25]. In the practice testing study involving problem solving [15] a different outcome may result if interactive multimedia elements were used to replace the paper-based problems.

3 Practice Testing – Design Approach

When designing educational e-assessments it is essential to have supporting evidence to demonstrate the reliability and validity of the approach taken. Such evidence typically includes some form of domain analysis and mapping of learning outcomes to test specifications and blueprints [12]. A wide range of materials as described earlier have been employed in the practice testing field. The development and selection of future materials should be considered within a formal assessment development framework and mapped to curricular learning outcomes accessing a wider range of thinking skills supported by item banks.

3.1 Item Bank Development

One of the key components of online assessment is a well-structured item bank where the items have gone through a systematic process of authoring, reviewing and testing before implementation. A test item is a question or task that requires a response from a student and is subject to a scoring rule. Test items are divided broadly into two types of student response; constructed response or selected response. In the assessment literature, considerable attention has been directed at the limitations of selected response items and in particular multiple-choice items in testing higher-order cognitive abilities. The use of tests with multiple-choice items to assess students is widespread in both formative and summative online testing. While multiple-choice and True/False items are the most well-known of selected response items a range of item types are supported for online delivery under the IMS Question & Test Interoperability (QTI) [17] specification.

3.2 Online Practice Testing Considerations

The IMS QTI specification describes a data model for the representation of question and test data and enables the exchange of items and results data. This model is illustrated in Figure 3 and identifies in a simplified way the roles of the various actors involved.

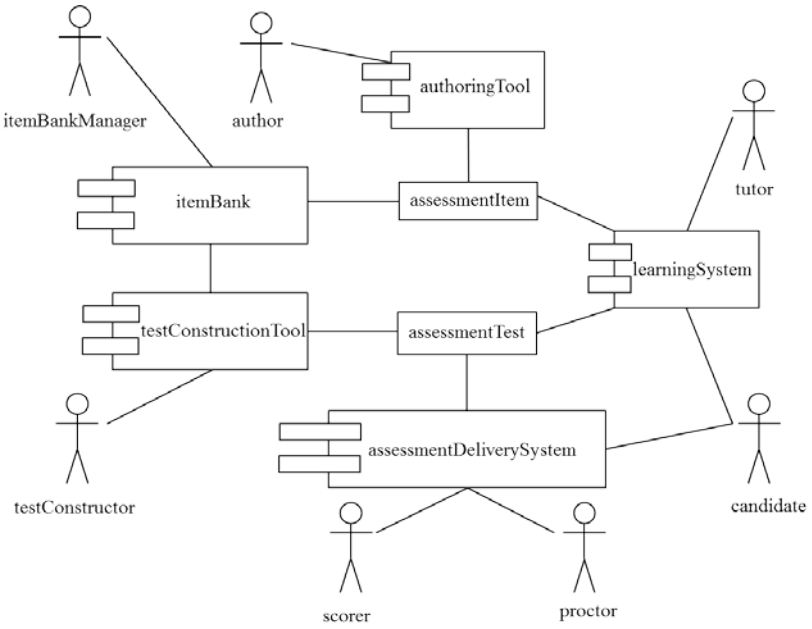


Figure 3: The Role of Assessment Tests and Assessment Items [17]

The Test author can create items and build an item bank using an authoring tool. Moodle provides an item authoring interface, or an alternative approach is to use offline QTI compliant authoring tool such as Respondus 4.0 [28] which can create, manage and publish online tests or quizzes to a number of supported LMS's. Test items can be imported into Respondus from suitable file formats such as MS Word or Excel which packages the test items for publishing to different LMS personalities or to a generic QTI file format. The major benefit of using a tool such as Respondus is the ability to publish a test to multiple courses taking the same module (requires a Campus Wide Licence). Tests are published directly to the course page and are constructed from item banks in accordance with the test specification or blueprints [12]. The student is represented by the candidate actor in the QTI assessment model. A student interacts with the practice test typically through an LMS such as Moodle. A number or user interactions are supported by the QTI specifications including list based interactions, graphical interactions, text interactions and miscellaneous interactions. These interactions facilitate a range of item types and cognitive levels [2].

Current trends in education include a range of emerging technologies such as flipped classrooms and massive open online courses (MOOC's). In both of these test items are embedded in the online material. In a flipped classroom, students typically watch video lectures online with quizzes included with the video material before coming to class. The face to face class time is spent on activities that reinforce their learning. In the MOOC, video material is typically presented with simple assessment activities such as embedded MCQ and True/False items. The research to extend practice testing to include higher order thinking skills such as problem solving will have potential applications in these emerging areas.

3.3 Problem Solving Environment

Problems vary considerably along a continuum ranging from well-structured to ill-structured problems. The type of problem solving suitable for online testing could extend to midway on the continuum proposed by [21] to include; algorithmic problems; story problems; rule-using problems; decision-making problems; troubleshooting and diagnosis-solution problems. Test items could be viewed as mini problem solving environments following an instructional model called problem-centred learning (PCL). PCL uses real world problem scenarios or tasks consisting of six components as illustrated in Table 2.

Table 2: Components of PCL adapted from [9]

Component	Description	Question Item
Trigger Event	the problem scenario presented to the student	question stem
Case Data	information about the scenario	multimedia or other element
Case Resolution	actions necessary to resolve problem	solution determined
Guidance	instructional resources to keep student on track	in question resources
Feedback	information in response to student actions	correctness of response
Reflection	student reflection	student reflection

The item stem is mapped to the trigger event where the problem scenario is described or presented. The case data provides more background information to the problem scenario which could be reflected as a multimedia element, graphic or table within the test item. The case resolution depending on the complexity of the problem scenario and when challenged to solve a problem students must draw on related knowledge and skills which can lead to "teachable moments" [8] and are not unlike "moments of contingency" in formative assessment providing critical points in instruction where learning can change direction depending on an assessment [3].

4 Conclusions and Recommendations for Future Work

The aim of this paper was to provide a survey/state of the art on practice testing by analysing the methodologies employed by the practice testing community. Practice testing has been shown to be beneficial for learning and online practice testing can be an effective technique to help students to regulate their learning. Assessment strategies need to shift away from end of learning summative assessment to more use of embedded practice testing where assessment is seen and used as an effective technique to enhance student learning. In order to inform this shift to embedded practice testing further research needs to be conducted using e-assessment to determine the boundaries of the testing effect in areas such as transfer of learning and higher order thinking. The challenge for technologists and educators alike is to leverage the potential of e-assessment for higher order learning and not simply replicate traditional paper-based tests in an electronic format.

References

[1] Anderson, L.W. and Krathwohl, D.R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*.

[2] Bacon, D. (2012). Use of Question Types and Features in QTI. *CAA 2012 International Conference, University of Southampton*, (2012), 1–13.

[3] Black, P. and Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation & Accountability*. 21, 1 (2009), 5–31.

[4] Bloom, B.S. (1956). *Taxonomy of Educational Objectives*.

[5] Butler, A. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*. 13, 4 (Dec. 2007), 273–281.

[6] Butler, A. and Roediger, H.L.I. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*. 36, 3 (2008), 604–616 LA – English.

[7] Carpenter, S.K. (2012). Testing Enhances the Transfer of Learning. *Current Directions in Psychological Science*. 21, 5 (Oct. 2012), 279–283.

[8] Clark, R. (2008). *Building expertise: Cognitive methods for training and performance improvement*. Pfeiffer/John Wiley & Sons.

[9] Clark, R. and Lyons, C. (2010). *Graphics for learning: Proven guidelines for planning, designing, and evaluating visuals in training materials*. Pfeiffer/John Wiley & Sons.

[10] Cranney, J. et al. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*. 21, 6 (Sep. 2009), 919–940.

[11] Daniel, D.B. and Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*. 31, 3 (2004), 207–208.

[12] Downing, S.M. and Haladyna, T.M. (2006). *Handbook of Test Development*. Lawrence Erlbaum Associates, Publishers.

[13] Dunlosky, J. et al. (2013). Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology. *Psychological Science in the Public Interest*. 14, 1 (Jan. 2013), 4–58.

[14] European Commission (2011). COMMISSION REGULATION (EU) No 1149/2011. *Official Journal of the European Union*. 05 (2011), 1–124.

[15] Van Gog, T. and Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science*. 36, 8 (Nov. 2012), 1532–1541.

[16] Haladyna, T.M. (1997). *Writing Test Items to Evaluate Higher Order Thinking*. Allyn & Bacon.

[17] IMS QTI: (2013). *http://www.imsglobal.org/question/qtiv2p1/imsqti_oviewv2p1.html*. Accessed: 2013-10-11.

[18] Jensen, J.L. et al. (2014). Teaching to the test…or testing to teach: Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review*. (Jan. 2014).
[19] Johnson, B.C. and Kiviniemi, M.T. (2009). The Effect of Online Chapter Quizzes on Exam Performance in an Undergraduate Social Psychology Course. *Teaching of Psychology*. 36, 1 (Jan. 2009), 33–37.
[20] Johnson, C.I. and Mayer, R.E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*. 101, 3 (Aug. 2009), 621–629.
[21] Jonassen, D. (2000). Toward a design theory of problem solving. *Educational Technology Research & Development*. 48, 4 (Dec. 2000), 63–85.
[22] Kang, S.H.K. et al. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*. 19, 4/5 (Jul. 2007), 528–558.
[23] McDaniel, M.A. et al. (2011). Test-Enhanced Learning in a Middle School Science Classroom: The Effects of Quiz Frequency and Placement. *Journal of Educational Psychology*. 103, 2 (May 2011), 399–414.
[24] Morris, C.D. et al. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning & Verbal Behavior*. 16, 5 (Oct. 1977), 519–533.
[25] OECD (2013). *PISA 2012 Assessment and Analytical Framework. Mathematics, Reading, Science, Problem Solving and Financial Literacy*. OECD Publishing.
[26] Rawson, K. and Dunlosky, J. (2012). When Is Practice Testing Most Effective for Improving the Durability and Efficiency of Student Learning? *Educational Psychology Review*. 24, 3 (Sep. 2012), 419–435.
[27] Rawson, K.A. and Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*. 140, 3 (Aug. 2011), 283–302.
[28] Respondus 4.0: Exam Authoring Tool: *http://www.respondus.com/products/respondus/index.shtml*. Accessed: 2014-02-21.
[29] Roediger, H.L.I. et al. (2011). Test-Enhanced Learning in the Classroom: Long-Term Improvements from Quizzing. *Journal of Experimental Psychology: Applied*. 17, 4 (Dec. 2011), 382–395.
[30] Roediger, H.L.I. and Karpicke, J.D. (2006). The Power of Testing Memory Basic Research and Implications for Educational Practice. *Perspectives on Psychological Science (Wiley-Blackwell)*. 1, 3 (Sep. 2006), 181–210.
[31] Roediger, H.L.I. and Marsh, E.J. (2005). The Positive and Negative Consequences of Multiple-Choice Testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 31, 5 (Sep. 2005), 1155–1159.
[32] Rohrer, D. et al. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 36, 1 (Jan. 2010), 233–239.
[33] Simon, H.A. and Newell, A. (1971). Human problem solving: The state of the theory in 1970. *American Psychologist*. 26, 2 (Feb. 1971), 145–159.
[34] Slamecka, N.J. and Katsaiti, L.T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 14, 4 (Oct. 1988), 716–727.

Session 6

Green Computing

ChargeFlow - A workflow based solution to chargeback processing

Shankar V Subramanian[1] & Donna O'Shea[2]
[1]CS-Advanced Storage Division, EMC Software and Services, India
[2] Department of Computing, Cork Institute of Technology, Ireland
shankar.v.subramanian@emc.com; donna.oshea@cit.ie

Abstract

The focus of this paper will present *ChargeFlow* which is a workflow based approach to manage the issue of chargeback within a Cloud environment. This work has been driven by the need for both private and public cloud operators to manage their virtualized infrastructures to ensure proper accountability so that resource usage is metered, charged and billed to the respective user of the virtualized resource. ChargeFlow essentially automates and orchestrates the process of chargeback and charge view reducing the administrative overhead and increasing utilization of underlying resources.

Keywords: chargeback, chargeview, cloud, IaaS, accountability

1 Introduction

Cloud computing is the next big step in distributed computing. Its goal is to make better use of distributed resources and has many application areas, such as software testing. Testing software typically involves examining the behavior and operation of the software over multiple platforms and versions with various dependencies in multiple configuration settings. Before Cloud computing and virtualization became a reality this usually meant that the test team would have a number of resources on site and these resources had to be shared between various teams across multiple departments. A considerable amount of time was spent by the test team in getting the resource 'ready' for testing as it frequently involved installing various versions of an operating system on the resource to validate that the software functioned as designed.

With Cloud computing and virtualization, computing resources can now be created quickly with different operating systems and versions to test the software on various platforms. Within a private cloud these infrastructure services are offered to co-operate units or departments within a company, whereas public clouds provides infrastructure on demand services to third parties.

Managing these virtual environments is a very challenging issue. Cloud operators are offering essentially their Infrastructure as a Service (IaaS) and need to allocate the costs of IT resources to the people or organizations who consume them. Charging for usage will typically ensure that the virtual resources are used efficiently helping drive up utilization rates. Some type of chargeback tool is required to facilitate accountability and provide proper: resource usage metering; charging and billing; and reports detailing consumption and charges. Lack of a charging model and associated architecture typically causes problems such as resource 'hogging' where individuals try and retain the resources as long as possible even though they have no use for them. In addition, companies or operators availing of IaaS want to be clear in advance about how they are going to be charged and visibility to charges on a daily basis. According to [1] public cloud operators that have enterprise ready features relating to chargeback are likely to be more successful IaaS providers.

While the above is clearly important a recent study provided by [2] outlined that 49% of private cloud providers did not support any charging mechanism, with 24% of respondents stating that reporting was available with no charge been applied and only 22% having a full chargeback solution in place to the end user or department.

As a result of the above, it is the purpose of this paper to present *ChargeFlow* which is essentially a chargeback model and associated architecture that uses workflows to represent the procedural actions and decisions associated with chargeback. According to the Workflow Management Coalition a workflow is defined as "*the automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules*" [3]. According to [4] the ability to separate functions from applications is one of the biggest advantages to workflows. In addition, using a workflow based approach to chargeback means that the workflows themselves are easy to understand and modify.

The outline of this paper is as follows. Firstly, a state of the art with regard to existing chargeback solutions will be presented. Following this, the ChargeFlow architecture will be presented and details on how it was implemented and tested on VMware vCloud software suite. Summary and conclusions on the work will also be presented.

2 State of the art

There has been a number of works within the area of chargeback for IaaS cloud providers over the past number of years. These contributions have mainly focused on determining costs and charging for the resources managed in the cloud.

Yu et al [5] presents a pricing and charging model for the Cloud Simulation Platform which is a simulation and modelling platform for cloud services. The paper focuses on developing a usage based pricing strategy. Within this strategy there are three basic input parameters: basic cost; basis time and overtime rate, which are adjusted according to whether the resources are used for a short or long period of time. A billing module was also presented within the scope of this paper which manages the: accounts; price; and records relating to the simulations ran. However this work was focused on a simulation environment and not a physical test bed. ChargeView in [6] presented a means to automate the process of IT costing and chargeback. It identified cost attributes such as unit; batch; customer and facility level costs associated with cloud resources, collected usage data, calculated cost and reported this to the user. [7] developed an energy chargeback tool to report and expose service energy consumption and bill users accordingly for cloud resources. The rationale behind this work was to raise energy awareness among users and to foster a more sustainable infrastructure usage approach. The work in [8] presented a method for apportioning server costs among workloads in a shared resource environment such as the cloud. It represents the total cost of a resource pool including the acquisition costs for facilities, physical IT equipment, power and administration costs and defines the final cost for an application workload to be the sum of the costs over all costing intervals. The work focused on developing costing information which can potentially be fed into a pricing model, while [9] presented a cost-profit petri net to show the charging models and relationships between cost and profit. In addition, Armbrust et al in [10] proposed a tradeoff charging model and revenue equation for adverts supported model for cloud services.

Commercial contributions such as [i, ii, iii, iv] are also available for private cloud operators. Most of these commercial solutions however are dependent on other products within their suite of offerings and cannot be easily deployed within an existing private cloud or as a standalone product. In addition, such commercial tools are expensive and their functionality is generally limited on one type of resource (e.g. storage). None of these works to date, however, use a workflow based solution to chargeback processing, the approach of which is presented in this paper.

3 ChargeFlow Architecture

The driving force behind this work was to automate the existing chargeback workflow within the EMC software and services India IaaS group. Within this private cloud the existing chargeback model had significant issues where a script ran daily to determine active VMs was easily circumvented by users creating VMs in the morning and deleting them before the script ran. A more sophisticated workflow solution was required.

Generally workflows according to [v] consist of actions and decisions that you run sequentially and consist typically of schemas, attributes and parameters. The workflow schema is the main component of a workflow as it defines all the workflow elements and the logical connections between them while the workflow attributes and parameters are the variables that workflows use to transfer data. There are a number of workflow schemas defined in *ChargeFlow* to take into account the: create; delete; modify and chargeback operations that can be performed by the user/actor of the system. These workflows are further outlined in the implementation section. To support the workflows there are a number of architectural elements required which are categorized into layers i.e. provisioning; physical; collection and chargeback. These layers are described in Table 1 and the flows between the layers are further described in Figure 1.

[i] IBM Tivoli Usage and Accounting Manager, http://www.ibm.com/software/tivoli/products/usage-accounting/
[ii] HP Storage Essentials Chargeback Manager, http://h18006.www1.hp.com/products/storage/software/e-suite/wf05-chargeback.html
[iii] Northern Storage Chargeback, http://www.northern.net/nss
[iv] VMware vCenter Chargeback, http://www.vmware.com/products/vcenter-chargeback
[v] VMware vSphere Documentation Centre, Concepts of Workflows, http://pubs.vmware.com/vsphere-51/index.jsp?topic=%2Fcom.vmware.vsphere.vcenterhost.doc%2FGUID-2773B21B-8F4F-435E-BCC2-5833BFA82761.html

Layer	Description
Provisioning Layer	Within this layer there exists a portal that facilitates users in accessing resources within the infrastructure.
Physical Layer	The physical layer provides access to the underlying resources of the system and is the location where the ESXi hypervisors of the VMs are deployed.
Collect Layer	Contains the Orchestrator server and the *ChargeFlow* workflows for capturing the chargeback operations. It also contains the SQL server for updating the captured information.
Chargeback Layer	Front-End portal for administrators to initiate chargeback operation.

Table 1 Chargeflow layered architecture components

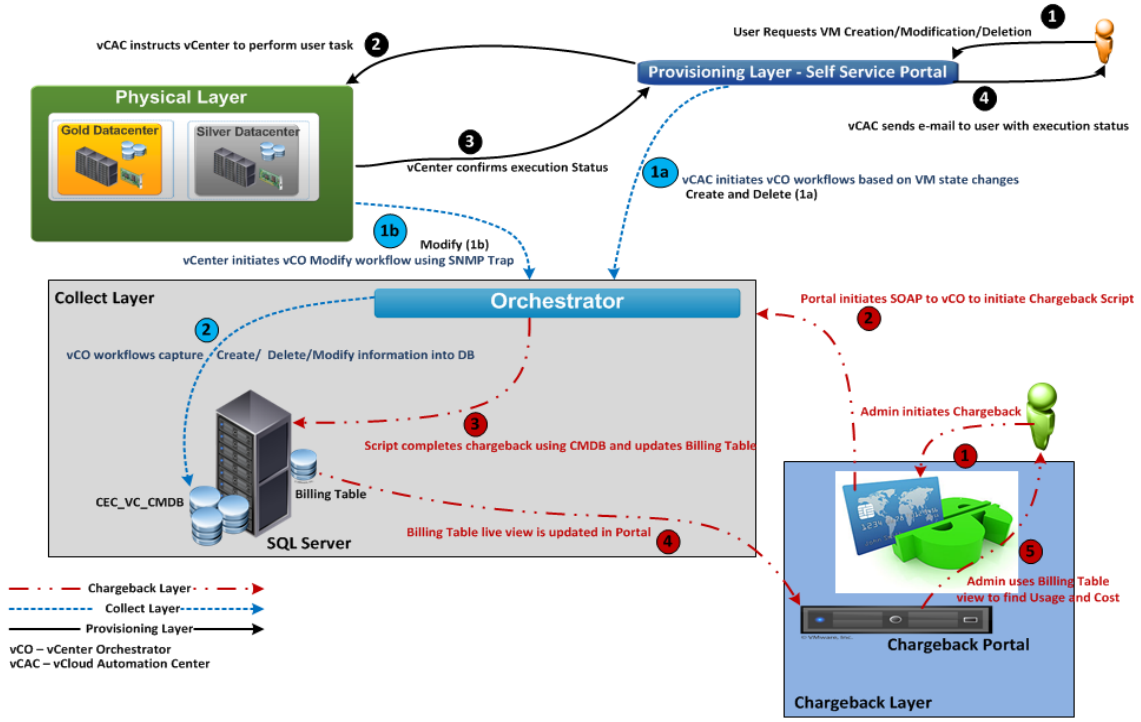


Figure 1 Workflow between Layers

To provide a common format for the charging and billing information within *ChargeFlow* an XML Schema Definition (XSD) was defined that specifies the rules to which the billing XML must conform in order to be considered valid within the *ChargeFlow* architecture. Within the system each user/department or company can have a different cost profile associated with them in accordance to their SLAs. Allowed profile types include: bronze, silver and gold with allowed resource types defined as: memory, storage and CPU. Depending on the profile and resource a cost is applied to the user/department or organization. This information is then feed into the chargeback process used in ChargeFlow the details of which are further outlined in the Section 4.

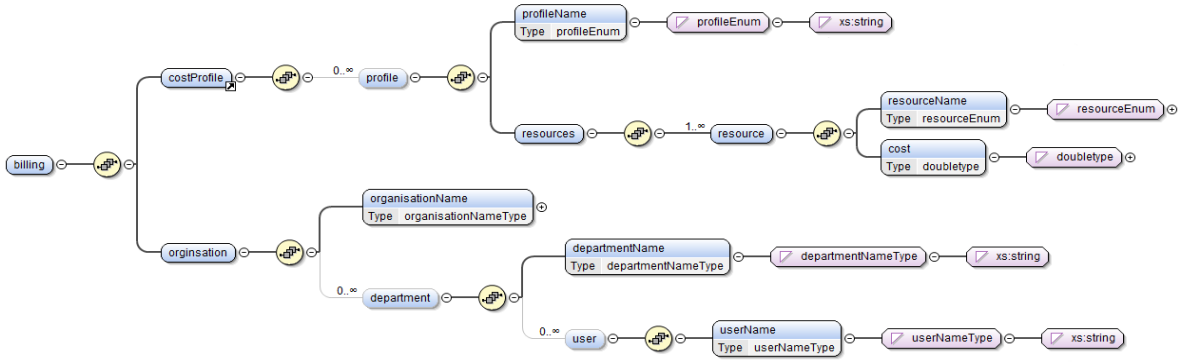


Figure 2: XSD for charging & billing

4 ChargeFlow Implementation

In order to provide a proof of concept for *ChargeFlow*, the workflows described in the previous section were implemented using software bundled within the VMware vCloud Suite [vi]. Within this suite of software various components were used to demonstrate *ChargeFlow*'s functionality, which are further outlined in Table 2. In addition to the components described below in Table 2, ChargeFlow also used Oracle's SQL Database to record the cost structure and changes that affect the charge associated with a user or department. ChargeFlow can also be implemented with open source systems such as Open Stack and the Nova (compute) layer will communicate with multiple virtualization technologies and help us to implement this concept in Open Source models.

Component	Tier	Function
vCAC (vCloud Automation Centre) Self-Service Portal	Presentation	Web portal that allows users to request resources from a catalogue of services.
vCAC (vCloud Automation Centre) Admin Console	Resource	Provisions the underlying infrastructure by using reservation pools.
vCenter Orchestrator	Orchestration	Orchestrates the provisioning of the underlying resources for a customer's Private Cloud.
vCenter Server	Physical	The virtual machines are deployed at this layer within a ESXi host. This tool is used to manage multiple ESXi hosts. The template of the Virtual Machine with the Operating system image is used to clone and deploy the Virtual Machine on an available ESXi host. The templates are chosen based on the Blueprints selected by customer on the vCAC Self Service Portal.

Table 2 Components Tier and Functions

4.1 ChargeFlow Process

Using the system components described above the following section will describe the workflows used to facilitate the processes associated with *ChargeFlow*.

```
<?xml version="1.0" encoding="UTF-8"?>
<billing
xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:noNamespaceSchemaLocation="file:/C:
Chargeflow/src/CostTempalte.xsd">
    <costProfile>
        <profile>
            <profileName>Bronze</profileName>
            <resources>
                <resource>
<resourceName>Memory</resourceName>
                    <cost>19</cost>
                </resource>
                <resource>
<resourceName>CPU</resourceName>
                    <cost>113</cost>
                </resource>
                <resource>
<resourceName>Storage</resourceName>
                    <cost>0.94</cost>
                </resource>
            </resources>
        </profile>
    </costProfile>
    <orginsation>
        <organisationName>CIT</organisationName>
        <department>
<departmentName>Computing</departmentName>
            <user>
                <username>Donna O'Shea</ sername>
            </user>
        </department>
    </orginsation>
</billing>
```

Table 3 Sample XML billing data

4.1.1 VM Creation

To begin, the user needs to login to the vCloud Automation Centre (vCAC) Self Service portal which resides at the presentation tier of the *ChargeFlow* architecture. This portal is used to initiate VM creation using the pre-configured BluePrints (BPs). BluePrints (BPs) are fundamental building blocks for provisioning a VM and represent processes and policies that can be applied to a VM [vii]. Using BPs essentially allows the users to choose their resources from a minimum to a maximum allocation level as specified within their SLA. A typical default blueprint creates a VM with: 2 CPUs; 4G of memory; and 40GB storage. There is a static cost associated with a blueprint which is typically between $ 0.930411 - $3.671233 per day or $0.03 - $0.15 per hour. This cost is the default cost with the real cost value been extracted from the users billing data stored in the database in XML form. A sample of this XML is shown in Table 3.

Once the VM has been created *ChargeFlow* monitors the state of the VMs through its custom orchestration workflows associated with the: create; modify and delete operations that are allowed on the system by the user.

[vi] Data Sheet VMware vCloud Suite Standard, Advanced and Enterprise Editions http://www.vmware.com/products/vcloud-suite
[vii] vCloud Automation Center Operating Guide for vCloud Automation Center 5.1, http://www.vmware.com/documents

These workflows are outlined in more detail in the next sections.

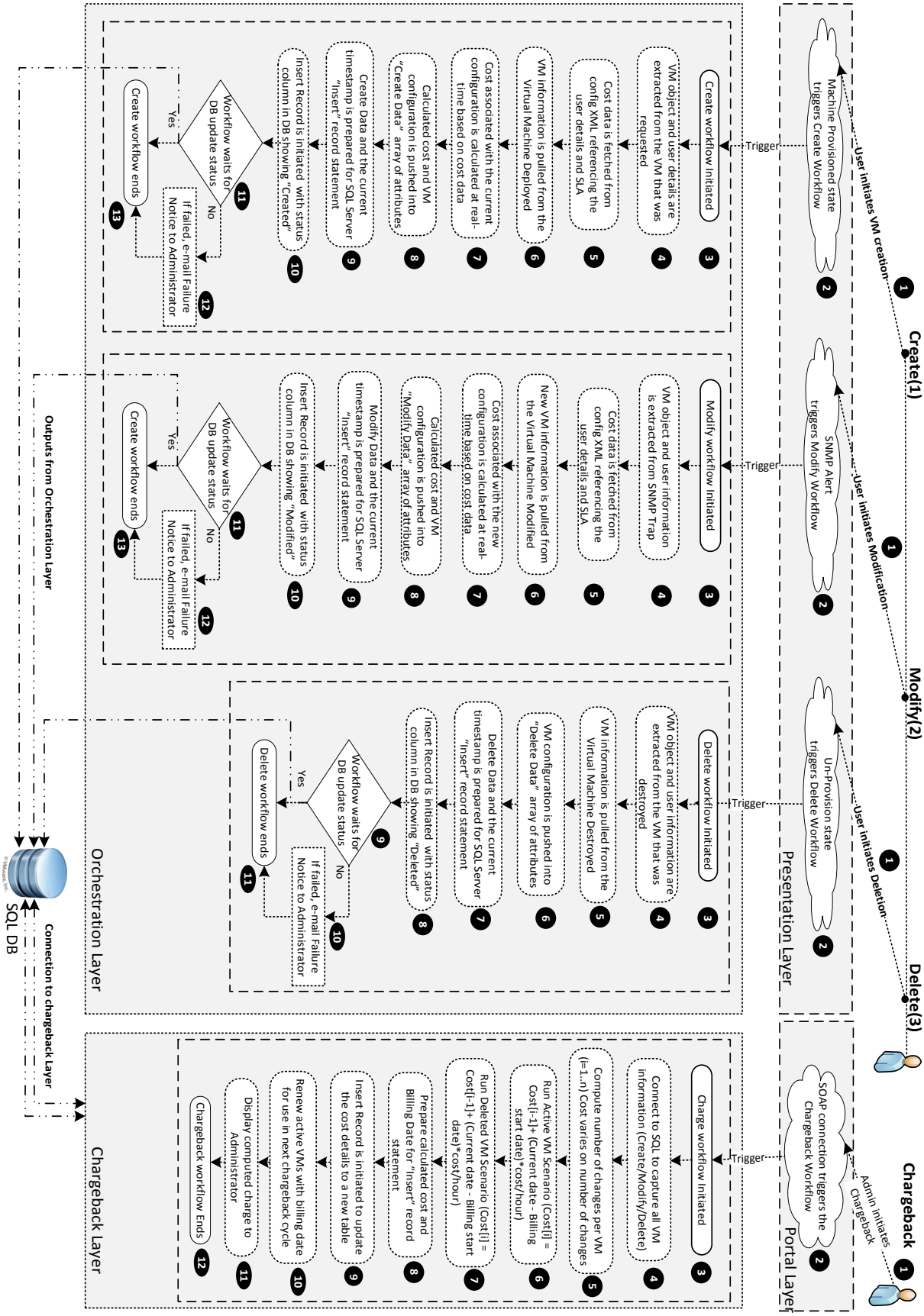


Figure 3 ChargeFlow custom orchestration workflows

4.1.2 Create Workflow

Virtual Machines are requested by users from the vCAC self-service portal. The VM creation passes state changes like "*Requested, Building Machine, Machine Provisioned, TurningOn and On*" before they are available for users.

When the VM is created, *ChargeFlow* ensures vCAC server to monitor for the state *Machine Provisioned.* Once the state occurs it triggers the execution of the custom orchestration workflow we refer to as the create workflow within *ChargeFlow*. The create workflow has the responsibility of collecting information such as: VM name; VMUUID (unique ID associated with VM); Owner; OS; IP Address; CPUs; Memory; disk space; server tier; status; timestamp; and date. The cost data is returned from the database in XML form based on the user ID of the person who created the VM. In the case where there is no cost data for the user then the user will be mapped to a department and the cost data associated with the department will be returned. The custom orchestration create workflow is outlined in detail in the Figure 3[Create (1)] provided below.

4.1.3 Modify Workflow

In the case where the user wants to modify the VM to include more or less resources then the charge been applied to them also needs to be modified. The modify custom orchestration workflow manages this operation. As vCAC does not offer a state change for VM modification, this workflow is triggered through an SNMP trap generated by the vCenter Server at the physical layer and inserts a new record with the modified resource details such as: CPU; memory; disk and cost in addition to the timestamp in which the VM resource was modified. This Modify workflow is outlined in detail in the Figure 3 [Modify (2)] provided below.

4.1.4 Delete Workflow

In the case where a VM is deleted, the VM deletion passes through state changes like "*Deactivate machine, Unprovision Machine and Machine Disposing*". *ChargeFlow* ensures that vCAC monitors for the state change *Unprovision Machine* and triggers the Delete workflow within the orchestration layer. This custom workflow collects the same information as the create workflow but excludes the step to calculate the cost. This Delete workflow is outlined in detail in the Figure 3 [Delete (3)] provided below.

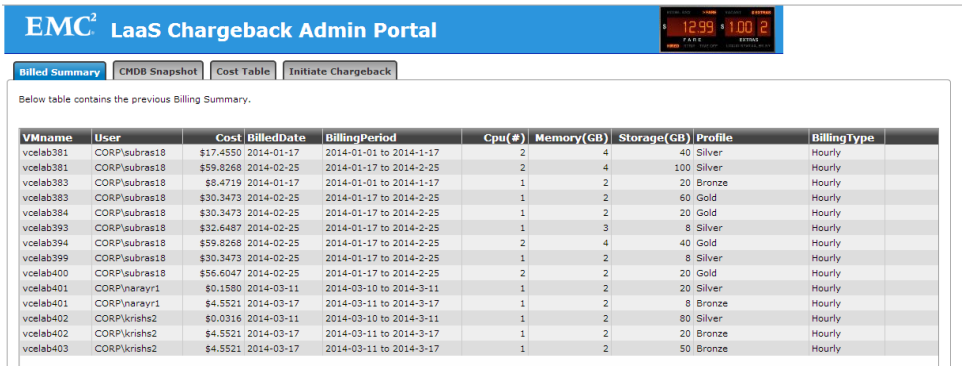


Figure 4 Chargeview web portal

4.1.5 ChargeFlow chargeback and charge view

An essential element of ChargeFlow is the ability for users/managers to be able to login to the system to view the charges been applied to them on a daily basis. This workflow resides at the orchestration layer of the architecture and generates the data necessary to compute the charge been applied.

The user can initiate this workflow by logging into the *ChargeFlow* chargeview web portal designed for system administrators and managers. A screenshot of this web portal is shown in Figure 4 above. On logging in, the portal establishes a Simple Object Access Protocol (SOAP) connection to the orchestration layer of the architecture and executes the workflow associated with calculating the chargeback, which is outlined in more detail in the Figure 3. *Chargeflow's* chargeback process is calculated based on the time period of usage also known as the billing period and is automated to calculate the chargeback between the last billed data and the current date. Between these dates active and deleted VMs used during the billing period are returned from the database and based on this information the following scenarios are considered.

4.1.5.1 Active VMs

Within the billing period it is possible that an active VM can either be: created; modified or renewed. The last date is taken as the current date to calculate the number of days and hours the VM has been active. The workflow calculates the time-period (i.e. the time difference between each change of configurations) for each configuration of the VM and then multiplies the configuration cost in each row and calculates cost accordingly, as described in Table 4. Each row would have a cost associated to it, that cost is accounted for that specific time-period.

Cost [I] = Cost [I - 1] + (Current date - Billing start date or VM create date in hours) * cost per hour Where, I is the number of changes 1...n.

Table 4 Chargeback calculation for active VMs

Once the cost is calculated the Active VMs are updated with a status "Renewed" and with the current timestamp of chargeback calculation. This facilitates the next billing cycle to calculate from the previous billing date and time.

4.1.5.2 Deleted VMs

If a VM is deleted, the difference between create timestamp and delete timestamp is considered for chargeback processing. The calculation procedure is described below in Table 5. As the deleted VMs are not required to be renewed, they are archived into another table for future reference. This ensures the CMDB always has a VM which is active or renewed.

Cost [I] = Cost [I - 1] + (Current date - Billing start date or VM create date in hours) * cost per hour Where, I is the number of changes 1...n

Table 5 Chargeback calculation for deleted VMs

5 Conclusion and Future work

This paper presented the use of workflows to manage the activities of chargeback and chargeview for Cloud providers offering IaaS. Within this paper the ChargeFlow architecture and implementation workflows were described in detail and a proof of concept was presented on a physical test bed using the VMware vCloud Solution. Within this work a static costing mechanism was used to represent the cost of a resource at a user/department/company level. Future work involves taking the architecture and workflows and incorporate a marketplace where resources can be priced based on the supply and demand on the resource been used. Further enhancements involve taking the idea of ChargeFlow to monitor the usage and energy consumption of resources facilitating green IT management. ChargeFlow can also be enhanced to work with open source systems such as OpenStack which can facilitate the usage of open source along with multiple virtualization technologies.

6 References

[1] B. Cohen, PaaS: *New opportunities for Cloud Application Development*, IEEE Computer Vol46, Issue9

[2] A. Wittmann, *InformationWeek 2014 Private Cloud Survey*, reports.informationweek.com

[3] WFMC (1996) *The Workflow Management Coalition Specification*, Terminology & Glossary, Issue 2.0 http://www.aiai.ed.ac.uk/project/wfmc/ARCHIVE/DOCS/glossary/glossary.html

[4] X. Liu, D. Yuan, G. Zhang, W. Li, D. Cao, Q. He, J. Chen, Y. Yan (2012) *The Design of Cloud Workflow Systems*, Springer Briefs in Computer Science.

[5] L. Yu, H. Zhang, (2012) *An effective Monetization Model and Pricing and Charging Strategy on Cloud Simulation Platform*, Proceedings IEEE 16th International Conference on Computer Supported Cooperative work in Design.

[6] S. Agarwala, R. Routray, S. Uttamchanani (2008) *ChargeView: An Integrated Tool for Implementing Chargeback in IT Systems*, IEEE Network Operations and Management Symposium, Salvador, Bahia,

[7] M. Gómez, D. Perales, E. Torres (2009) An Energy-Aware Design and Reporting Tool for On-Demand Service Infrastructures, *IEEE/ACM International Conference on Grid Computing*.

[8] D. Gmach, J. Rolia, L. Chrekasova (2011) *Chargeback Model for Resource Pools in the Cloud*, International Symposium on Integrated Network Management

[9] L. Bai, T. Li, X. Wu, Z. Xie (2011) *Charging Model Research of Infrastructure Layer in Cloud Computing based on Cost-project Petri net*, IEEE International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)

[10] M. Armbrust, A. Fox, R. Griffith (2009) *Above the clouds: a Berkeley view of cloud computing*, UC Berkeley Reliable Adaptive Distributed Systems Laboratory http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf

The Impact of Dynamic Monitoring Interval Adjustment on Power Consumption in Virtualized Data Centers

Mark White, Hugh Melvin, Michael Schukat

Discipline of Information Technology, NUI Galway, Galway, Ireland

m.white1, hugh.melvin, michael.schukat{@nuigalway.ie}

Abstract

Virtualization is one of the principle Data Center (DC) technologies increasingly deployed in recent years to meet the challenges of escalating costs, industry standards and the search for a competitive edge. In this paper we describe a novel approach to management of the virtualized system which dynamically adjusts the monitoring interval with respect to the average CPU utilization for the DC. We propose this will facilitate identification of cost opportunities which may otherwise have remained hidden mid-interval. We explain our adjustment algorithm and outline the architecture of the CloudSim test-bed with which we propose to continue evaluation of our hypothesis.

Keywords: Data Center, Energy Efficiency, Virtualization, Monitoring Interval Adjustment, CloudSim

1 Introduction

1.1 Virtual Machine Migration

In a virtualized data center, multiple Virtual Machines (VMs) are typically co-located on a single physical server, sharing the processing capacity of the server's CPU between them. When, for example, increased demands on the CPU result in reduced performance of one of the VMs to the point where a Service Level Agreement (SLA) may be violated, virtualization technology facilitates a migration. Migration relocates the services being provided by the VM on this 'over-utilized' host to a similar VM on another physical server, where sufficient capacity (e.g. CPU) is available to maintain SLA performance [1].

Conversely, reduced demand on the CPU of a host introduces opportunities for server consolidation, the objective of which is to minimize the number of operational servers consuming power. The remaining VMs on an 'under-utilized' host are migrated so that the host can be switched off, saving power. With power and cooling costs typically ranging from 30-45% of a DC's annual Operating Expenditure [2], server consolidation provides significant energy efficiency opportunities.

There are other reasons why a migration may be initiated in a data center (e.g. maintenance, Disaster Recovery (DR), hot-spot identification). Additionally, a range of metrics such as bandwidth, RAM and uptime are available to establish the performance of the computing infrastructure. We focus this paper on interval-based monitoring of CPU utilization. The average CPU utilization for the data center is the variable upon which we dynamically adjust the monitoring interval. Migrations are initiated if power consumption opportunities are identified.

1.2 CPU Utilization

In a virtualized environment CPU utilization is an indication of the processing capacity being used by a host while serving the requirements of the VMs located on it. In current practice, the CPU utilization

value delivered to monitoring systems is averaged over a constant monitoring interval. This interval is typically pre-configured (via a management interface) by the data center operator [3], rendering it static. Because a relatively small percentage of the host's CPU is concerned with running the virtualization hypervisor, CPU utilization is primarily dependent on the workload being serviced by the VMs located on the host. This workload varies with time. As such, the frequency of change of the CPU utilization value closely tracks the frequency of change of the incoming workload. CPU utilization, therefore, provides the most appropriate variable on which to adjust a dynamic monitoring interval.

1.3 Contribution

By dynamically adjusting the monitoring interval with respect to the incoming workload (indicated by a weighted CPU utilization average for the data center), we propose that there is potential for reduced power consumption through identification of performance issues at an earlier stage than systems which use a static interval.

The remainder of this paper is organized as follows: Section 2 describes some recent work related to our research. Section 3 introduces the CloudSim simulation software with which we will test and validate our algorithms. Section 4 details the design and initial implementation of our experiments. Section 5 presents a progress report on implementation. Section 6 concludes the paper.

2 Related Work

Voorsluys et al. [4] evaluate the cost of live migration, demonstrating that DC power consumption can be reduced if there is a reduction in migrations. The cost of a migration is dependent on a number of factors, including the amount of RAM being used by the source VM (which needs to be transferred to the destination) and the bandwidth available for the migration. The higher the bandwidth the faster data can be transferred. Additionally, power consumption is increased because 2 VMs (source and destination) are running concurrently for much of the migration process. A static monitoring window of 5 minutes is maintained throughout the Voorsluys experiments.

In their hotspot identification paper, Xu and Sekiya [5] select a monitoring interval of 2 minutes. The interval is chosen on the basis of balancing the cost of the additional processing required against the benefit of performing the migration. The 2 minute interval remains constant during experimentation.

Using an extended version of the First Fit Decreasing algorithm, Takeda et al. [6] are motivated by consolidation of servers, to save power. They use a static 60 second monitoring interval for their work.

Xu and Chen et al.[7] monitor the usage levels of a variety of server resources (CPU, memory, and bandwidth), polling metrics as often as they become available. Their results show that monitoring at such a granular level may not only lead to excessive data processing but the added volume of network monitoring traffic (between multiple hosts and the monitoring system) may also be disproportionate to the accuracy required.

The processing requirements of DC hosts vary as the workload varies and are not known until they arrive at the VM, requesting service. While some a priori analysis of the workload may be performed to predict future demand, as in the work of Gmach et al. [8], unexpected changes may occur which have not been established by any previously identified patterns. A more dynamic solution is required which reacts in real-time to the incoming workload rather than making migration decisions based on a priori analysis.

VMware vSphere facilitates a combination of collection intervals and levels [9]. The interval is the time between data collection points and the level determines which metrics are collected at each interval. Examples of vSphere metrics are as follows:

- Collection Interval: 1 day
- Collection Frequency: 5 minutes (static)
- Level 1 data: 'cpuentitlement', 'totalmhz', 'usage', 'usagemhz'

- Level 2 data: 'idle', 'reservedCapacity' + all of Level 1 data (above)

VMware intervals and levels in a DC are adjusted manually by the operator as circumstances require. Once chosen, they remain constant until the operator re-configures them. Manual adjustment decisions, which rely heavily on the experience and knowledge of the operator, may not prove as accurate and consistent over time as an informed, dynamically adjusted system.

In vSphere, the minimum collection frequency available is 5 minutes. Real-time data is summarized at each interval and later aggregated for more permanent storage and analysis.

3 Background

3.1 Over-Utilized Hosts in the CloudSim Framework

We performed our experiments using the CloudSim framework [10]. The default monitoring interval in a CloudSim simulation is 300 seconds. This reflects current industry practice where an average CPU utilization value for each host is polled every 5 minutes by virtualization monitoring systems (e.g. VMware). Identification of over-utilized hosts (which may result in SLA violations) is performed at each interval, resulting in migrations.

As the workload on a host increases, the host CPU becomes busier, moving towards the possibility of VM SLA violations [11]. The default CPU utilization threshold for an over-utilized host in CloudSim is 100%. An adjustable safety parameter is also provided by the default CloudSim framework, effectively acting as overhead provision. As an example, if the predicted CPU utilization value were 90% and was then multiplied by a safety parameter of 1.2, the resulting value of 108% would exceed the threshold. A safety parameter of 1.1 would result in a final value of 99% (for the same initial utilization), not exceeding the threshold. Neither the default upper threshold of 100% nor the default safety parameter of 1.2 was adjusted in our experiments.

Beloglazov et al. [12] conclude from their CloudSim experiments that Local Regression/Minimum Migration Time (LR/MMT) is one of the better algorithmic combinations to maintain optimal performance and maximize energy efficiency. We chose the LR/MMT algorithms as the basis for our own evaluation.

Having passed the most recent CPU utilization values through the LR algorithm, hosts are considered over-utilized if the next predicted utilization value exceeds the threshold of 100%. LR predicts this value using a sliding window, each new value being added at each subsequent interval throughout the simulation. The size of the sliding window is 10. Until initial filling of the window has taken place (i.e. 10 intervals have elapsed since the simulation began), CloudSim relies on a 'fallback' algorithm which considers a host to be over-utilized if its CPU utilization exceeds 70%. Our adjusted interval results in the 'first fill' of the window occurring sooner than the CloudSim default i.e. the longest interval (5 minutes) in our dynamic version is the minimum (only) interval in the CloudSim default.

VMs are chosen for migration according to MMT i.e. the VM with the lowest predicted migration time will be selected for migration to another host. Migration time is based on the amount of RAM being used by the VM. The VM using the least RAM will be chosen as the primary candidate for migration, simulating minimization of the Dirty Page Rate (DPR) during VM transfer [1]. The destination host for the migration is chosen on the basis of power consumption following migration i.e. the host with the lowest power consumption (post migration) is chosen as the primary destination candidate. In some cases, more than one VM may require migration to reduce the host's utilization below the threshold. Dynamic RAM adjustment is not facilitated in CloudSim as the simulation proceeds. Rather, RAM values are read (during execution of the MMT algorithm) on the basis of the initial allocation to each VM at the start of the simulation.

4 Interval Adjustment

Our dynamic interval adjustment algorithm involves two principle steps:

Table 1: Application of Weights to Predicted Utilization

Predicted Utilization (x_i)	**Weight Applied** (w_i)
1 - 10	1
11 - 20	2
21 - 30	3
...	...
91 - 100	10

1. Calculate the weighted mean of the CPU utilization value for all operational hosts in the data center as in Eq. 1. Non-operational hosts do not affect the result of the calculation.

2. Choose and set the next monitoring interval with respect to the weighted mean from Table 1 above.

$$x = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}, \tag{1}$$

where w_i is the weight applied to the range within which the predicted utilization value (x_i) for each operational host falls and n is the number of operational hosts.

We apply weights to the CPU utilization for each host so that host's which have a higher CPU utilization are given priority in the calculation. The monitoring intervals we apply to the resulting weighted average prediction are depicted in Fig. 1. We align our maximum interval with the existing interval in CloudSim i.e. 300 seconds. A minimum interval of 30 seconds facilitates 10 intervals in total, each having a corresponding 10% CPU utilization range from 0 to 100. It should be noted that the intervals and CPU ranges were chosen somewhat arbitrarily. They may well benefit from further analysis when the test-bed has been validated.

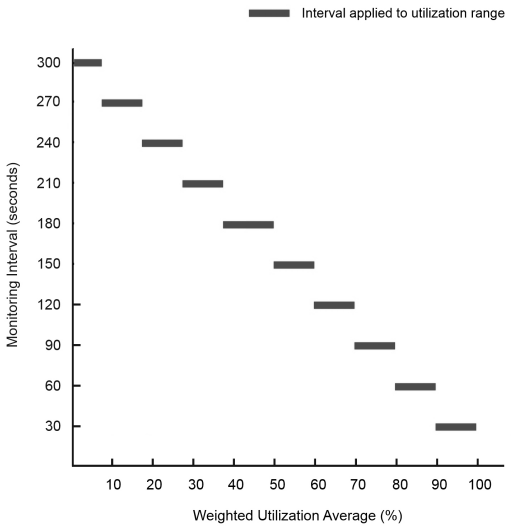


Figure 1: Calculation of the Monitoring Interval

5 Testbed Status

5.1 'Planetlab' Workload

By default, CPU utilization values in CloudSim are stored in 1,052 'Planetlab' text files. The values represent CPU utilization samples from 'real world' servers gathered from over 500 locations worldwide. The files, referred to as 'cloudlets' in CloudSim, simulate the workload to be processed by each VM in the data center. There are 288 CPU utilization values in each file, sufficient to accommodate reading a value every 300 seconds for the duration of the default CloudSim simulation i.e. 24 hours.

However, if our minimum interval of 30 seconds were applied for the full 24 hour simulation, 2880 values would be required (i.e. $(60 \times 60 \times 24)/30$) in each file. The default files provided by CloudSim are thus concatenated to ensure sufficient values are available for this scenario, resulting in a total of 105 complete files with 2880 values each.

Our experiments will focus on a comparison of the default CloudSim simulation with our dynamic version and must, therefore, use comparable workloads. Accordingly, we have created a new set of files for the default simulation. The values for these files have been calculated based on the average of the values used in our dynamic simulation by the time each 300 second interval had elapsed, as depicted in Fig. 2. A difference of less than 1% in the per-interval average CPU utilization of the workloads was observed thus validating the approach.

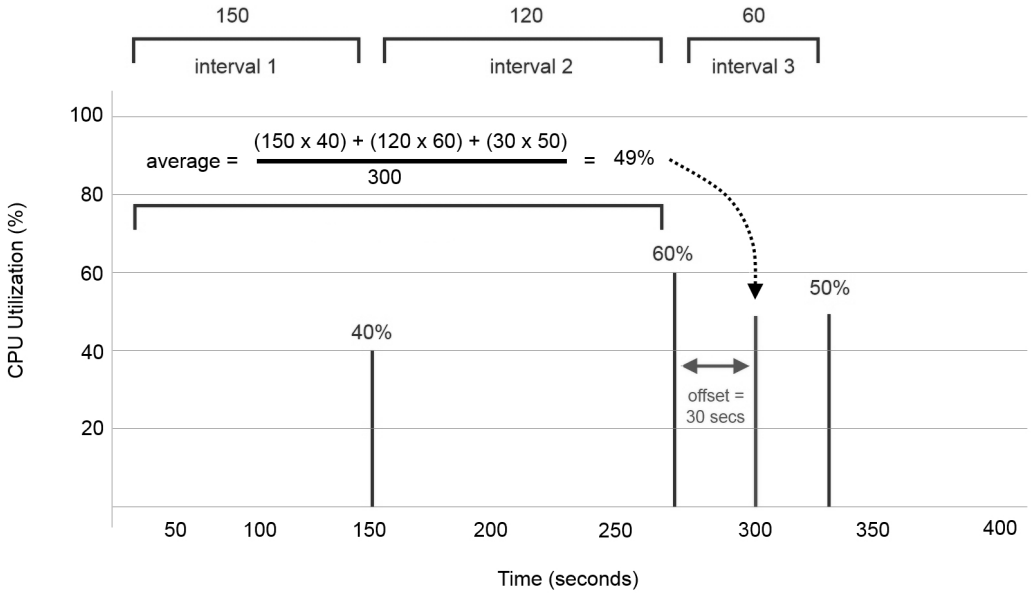


Figure 2: Calculation of the average CPU utilization for the default files

Additionally, as a result of the reduced file count after concatenation, the number of hosts running in the simulation has been reduced, from the default of 800 to 80, to maintain the ratio of VMs to hosts.

CloudSim workload is defined in MIPS (Million Instructions per Second). The values passed from the 'Planetlab' files are interpolated i.e. converted from CPU percentages to an equivalent MIPS representation. The resultant MIPS value is the workload that will be applied to each VM for the simulation. Each type of VM used in the CloudSim framework is assigned a MIPS capacity (e.g. 500, 1000, 2000, 2500) when the simulation begins. Likewise each CPU has an initial MIPS capacity of either 1860 or 2660. These configuration settings limit the number of VMs which can be run on each host and also the volume of workload which can be applied to each VM.

5.2 Testbed Validation Phase

In order to validate our approach, we have performed some preliminary tests based on comparison of the default LR/MMT CloudSim simulation with a simulation which includes our dynamic interval adjustment algorithm. The following configuration settings are common to both:

- Duration: 24 hours (86400 seconds)
- Available Hosts: 80
- Cloudlets (workload): 105 'PlanetLab' files
- Over-utilization threshold: 100%
- Safety Parameter: 1.2

Figure 3 depicts the intervals (727) calculated by our dynamic adjustment algorithm during a 24-hour simulation. The interval may potentially range between a minimum of 30 and a maximum of 300.

Initial test results suggest a reduction of 2-3kWh in power consumption. However, a reduced number of migrations is also reported by CloudSim. This is counter-intuitive. To complete validation of our testbed we are currently carrying out a full review of the CloudSim code.

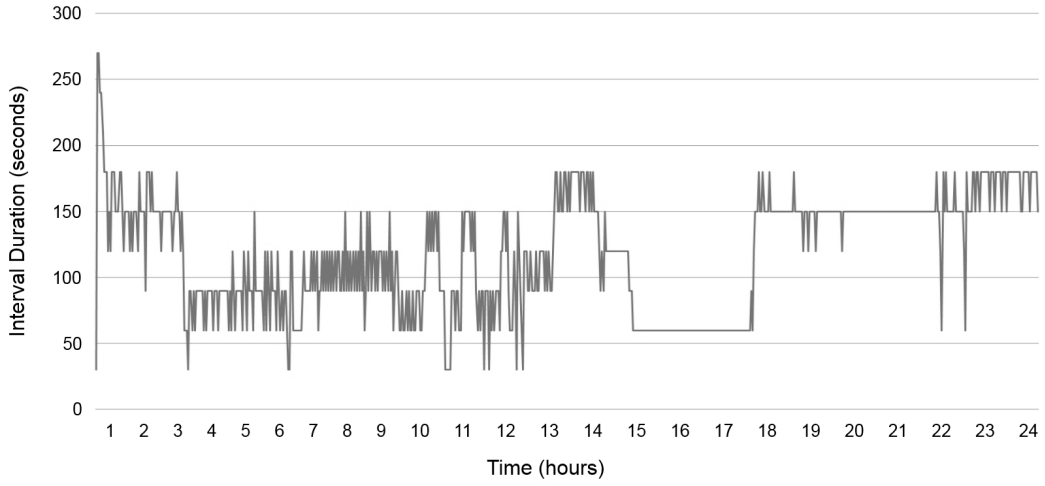


Figure 3: Intervals Calculated by the Adjustment Algorithm

6 Conclusions and Future Work

Our dynamic simulation is differentiated from the LR/MMT CloudSim in that we adjust the duration of the next interval (with respect to the weighted average of the data center's CPU utilization) rather than maintaining a static interval of 300 seconds. We aim to determine the impact of this approach on power consumption. Using our dynamic interval adjustment algorithm, analysis of CPU utilization is executed more often. Further investigation of the counter-intuitive migration results is being performed.

In our current work we focus on the over-utilization threshold. In calculating the average CPU utilization for the DC we apply shorter intervals as the average utilization rate increases. However, the adjusted interval should also reflect the urgency with which utilization should be monitored as the average approaches some *under-utilization* threshold. We intend to investigate this lower utilization threshold in future work.

References

[1] Clark, C., Fraser, K., Hand, S., Hansen, J. G., Jul, E., Limpach, Pratt, I., Warfield, A.: Live Migration of Virtual Machines. In: Proceedings of the 2nd Conference on Symposium on Networked Systems Design & Implementation, vol .2, pp. 273-286. USENIX Association (2005)

[2] Pakbaznia, E., Massoud, P.: Minimizing Data Center Cooling and Server Power Costs. Proceedings of the 14th ACM/IEEE International Symposium on Low Power Electronics and Design. ACM (2009)

[3] Elmroth, E., Lars L.: Interfaces for Placement, Migration and Monitoring of Virtual Machines in Federated Clouds. Grid and Cooperative Computing, 2009. GCC'09. Eighth International Conference on. IEEE (2009)

[4] Voorsluys W., Broberg J., Venugopal S., Buyya R.: Cost of Virtual Machine Live Migration in Clouds: a Performance Evaluation. In: Proceedings of the 1st International Conference on Cloud Computing. Vol. 2009. Springer (2009)

[5] Xu, Y., Sekiya, Y.: Scheme of Resource Optimization using VM Migration for Federated Cloud. In: Proceedings of the Asia-Pacific Advanced Network, vol. 32, pp. 36-44. (2011)

[6] Takeda, S., and Toshinori T.: A Rank-Based VM Consolidation Method for Power Saving in Data Centers. IPSJ Online Transactions, vol. 3 pp. 88-96. J-STAGE (2010)

[7] Xu, L., Chen, W., Wang, Z., Yang, S.: Smart-DRS: A Strategy of Dynamic Resource Scheduling in Cloud Data Center. In: Cluster Computing Workshops (CLUSTER WORKSHOPS), IEEE International Conference on, pp. 120-127. IEEE (2012)

[8] Gmach, D., Rolia, J., Cherkasova, L., Kemper, A.: Workload Analysis and Demand Prediction of Enterprise Data Center Applications. In: Workload Characterization, 2007. IISWC 2007. IEEE 10th International Symposium on, pp. 171-180. IEEE (2007)

[9] VMware, `http://pubs.vmware.com/vsphere-4-esx-vcenter/index.jsp?topic=/com.vmware.vsphere.bsa.doc_40/vc_perfcharts_help/c_perfcharts_collection_intervals.html` [16th April 2014]

[10] Buyya, R., Ranjan, R., Calheiros, R. N.: Modeling and Simulation of Scalable Cloud Computing Environments and the CloudSim Toolkit: Challenges and Opportunities. In: High Performance Computing & Simulation, 2009. HPCS'09. International Conference on, pp. 1-11. IEEE (2009)

[11] Iyer, R. K., Rossetti, D.J.: A Measurement-Based Model for Workload Dependence of CPU Errors. Computers, IEEE Transactions on, vol. 100, no. 6 pp. 511-519. IEEE (1986)

[12] Beloglazov, A., Buyya, R.: Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers. Concurrency and Computation: Practice and Experience vol. 24, no. 13, pp. 1397-1420. Wiley Online Library (2012)

Session 7

Wireless and Mobile Networks

Mitigating the Effects of Degraded Network Performance Metrics on RTP-based Video Streaming using a Neural Network Based Handover Approach

Sean Hayes, Enda Fallon, Ronan Flynn, Niall Murray

Software Research Institute, Athlone Institute of Technology, Athlone, Co Westmeath

shayes@research.ait.ie, efallon@ait.ie, rflynn@ait.ie, nmurray@research.ait.ie

***Abstract*— The substantial growth in the usage of mobile devices with internet access has resulted in increased demand for Internet access over wireless networks. Today, wireless networks that employ different technologies are often integrated to create large network topologies. However, there are challenges associated with combining different wireless technologies together as part of a large network topology. One such challenge is the variation in performance such as packet delay, interference and loss. The effects of jitter and packet loss have been well documented in the literature. However, not much research has been carried out on video streaming performance when initiating vertical handovers in heterogeneous network environments. This paper presents an objective view on the performance of video streaming in a heterogeneous network topology when a handover algorithm is implemented that hands the stream off to a better quality connection. The end result of the video stream is then compared to the original video file. The results of these experiments illustrate that in conditions where the current network is of poor quality, handovers can make a large difference to the overall quality of the received video file if a handover mechanism is implemented properly.**

Keywords: Heterogeneous Networking; Media Independent Handover; Video Streaming

I. INTRODUCTION

The effects of jitter and noise interference upon the overall quality of video files streamed through wireless networks have been the subject of previous research [1] [2]. More recently, delay and signal-to-noise ratio (SNR) have also been considered as important factors for streaming media [3]. With the growing usage of mobile devices for media streaming, the challenge of maintaining a sufficient quality of service (QoS) level when the user switches networks due to a fall in the QoS level of their current network is a major problem [4].

The Media Independent Handover (MIH) standard was developed by the IEEE 802.21 working group to facilitate a handover process between two internet layer technologies and allow for mobility of wireless devices between networks [5]. The MIH framework was developed to communicate critical network metrics to higher layer protocols as a result of link performance metrics falling to inadequate levels or a complete loss of connection. Most handover algorithms in this approach purely consider the RSS of nearby networks when initiating handovers. The handover method proposed in this paper can potentially improve the quality of the resulting video stream if the metrics used to measure video performance of adjacent networks are carefully weighed and used as a basis for network selection. Previous work [6][7][8] has given insight into how the presence of weather conditions can have a negative effect on a wireless signal by increasing the noise present on the network link which in turn degrades overall performance in wireless and mobile networks. In particular for video streaming, there is a noticeable impact on the quality of the video that the end user receives compared to the original source file. An approach is needed that properly considers the metrics that affect video quality essential to good QoS in video streaming when initiating a handover.

The aim of this performance metrics based algorithm is to help output a received video file that is as close as possible to the original source file. In this paper, a heterogeneous network scenario consisting of a WiMax/LTE fusion is considered in which poor weather conditions prevail. This requires the use of a suitable modulation strategy to avoid errors caused by interference by weather phenomena, meaning a modulation strategy with good resistance to noise interference but with a low data-rate is used. A selection of test results from previous brute force weighting tests are used, in which it is illustrated what performance metrics are most important for the best possible video output in the wireless topology. We normalize the performance metrics to fall between 0 and 100 (where 0 = bad and 100 = good) and then apply the weights to these values.

These tests will be repeated four times, each with the quality of the back-up LTE link present on the network topology further degraded through a combination of traffic from other users and

interference caused by rain in the area. This means an increase in loss and delay as the distance between the user and transmitter increases. This makes handing over to the WiMax connections much more important for a good video output. The mean opinion score of the received video file shall serve as the final verdict on the effectiveness of a particular weight configuration.

This paper is organized as follows: Section II outlines related work in the area of video streaming quality metrics and ways to ensure good performance. A brief overview of the technologies relevant to the work presented is given in Section III. The experimental set-up used and the results obtained are detailed in Section IV followed by conclusions in Section V.

II. RELATED WORK

The effect of performance metrics such as jitter and delay on the quality of received streamed video has been the subject of previous research. The use of jitter and packet loss as a performance metric is investigated in [9] where it is also demonstrated that in the absence of jitter and loss, the playback of a received video file is near identical to the original file. The results in [9] showed that a high jitter value coupled with high packet loss results in a poor quality received video while low jitter and loss values lead to a near "perfect video". What is interesting to note in these results is that the overall quality of the video does not decrease as dramatically if high jitter and low loss is applied or alternatively high loss and low jitter. The conclusion in [9] is that jitter is just as important a metric as data loss in regards to the effect on received video quality. The importance of the service provider in regards to high quality video streaming is discussed in [10]. The data-rate required for a stream of high quality video at a sufficiently fast buffer speed can only be supplied by service providers who broadcast at high data-rates and implement mechanisms to ensure good quality from the beginning to end of the video. Another critical performance indicator is the power consumption of streaming video on the user's wireless device. The experiments described in [11] show that implementing high quality streaming of multimedia may result in a dramatic decrease in the battery power of a device, particularly smartphones. It is further discussed in [11] that the power consumption of modern smartphones is inadequate for today's demands, and furthermore that network load and video quality are both major factors in terms of power consumption. It is concluded in [11] that the most efficient way to stream media on smartphones is to take into account that the users on their smart-phone screens don't usually see (perceive) much difference between very good quality and good quality and that slightly reducing the quality of the video stream sent at the server and transmitting less bits is preferable to transmitting more bits and potentially losing packets.

A unique method to ensure high quality media delivery in trains is outlined in [12]. Users on trains travelling at high speeds can experience unpredictable degradation of signal strength and delay due to factors such as increasing distance from wireless transmitters and interference from windy conditions which can in turn lead to high packet loss. The solution outlined in [12] is the usage of a UDP-based file delivery protocol to deliver the video. Experiments are carried out on a WiMax transmitter covering a rail network in Taiwan and the results showed an improvement in overall successful byte transmissions (throughput) in comparison to the FTP delivery method originally used. An analysis in [13] into the effects of network congestion on video quality shows that the random discarding of packets when a defined congestion level is reached on a network can lead to a reduction in quality of the received video stream. Modifying the way in which packets are discarded can help improve the received video quality. The proposed solution is to implement an appropriate packet discarding pattern, if the total number of discarded packets is less than a defined percent of all packets being generated. A method of controlling the end-to-end congestion on a network through the introduction of a new transport protocol called scalable streaming video protocol (SSVP) is outlined in [14]. By means of appropriate adjustments on the time between sent packets, this can result in a smoother flow and in turn a reduction of loss on the link.

The effects of handover's on a video stream have also been considered. In [15], video streaming is used as a performance measurement for the proposed handover method for wireless LANs. A scheme based entirely on delay is presented in which increases in delay are used to predict when packet loss will occur. The success of this delay-based method through the improved quality in the received video quality compared to traditional handover methods. In [16] a scenario in which a handover scheme is designed exclusively for IPTV (Sky, UPC television over the internet for example) streams is described. The method considers network congestion as a metric, with jitter used as the basis to minimize loss on a packet switching network. A client has two connections set up with a server using two separate WLANS. The connection with the best jitter performance is used to help decide which network to handover to. The effect of user mobility on the QoS

of a video stream is examined in [17]. Two scenarios are examined, the first is when the mobile node is moving while connected to a network and the second is when it is in the handover process. The results of the scenarios outlined demonstrate that the QoS of the mobile user that moves inside its own cell, regardless of its distance to the information source and the strength of the signal, is similar to the QoS of a stationary user.

These experiments illustrate that performance metric degradation on a wireless network, such as increased jitter and delay, can have a negative effect on the overall performance of the link, an example being streamed video quality. Metrics such as delay and jitter are also considered to be important metrics to monitor when considering handovers with video streaming. These finding are reflected in the results of the experiments presented in this paper.

III. TECHNOLOGY OVERVIEW

A. Media Independent Handover

MIH was devised by the IEEE 802.21 working group [18] as a standard allowing for a vertical handover from one wireless technology to another. It allows for the communication of network critical metrics to Upper Layer Mobility Protocols (ULMP) and has support for several wireless technologies including LTE, Wi-Fi and WiMax. MIH consists of three services; an event service, information service and command service. MIH enables predictive resource allocation prior to handover. The MIH function is a method for receiving event notifications, which details events that would require a handover to be initiated such as the sudden loss of a signal or quality degradation leading to inadequate QoS. Once an event occurs, a command is then passed down from the media independent handover function (MIHF) of the access point to the end user device's function. Figure 1 illustrates the interaction between the MIHF and other protocol layers.

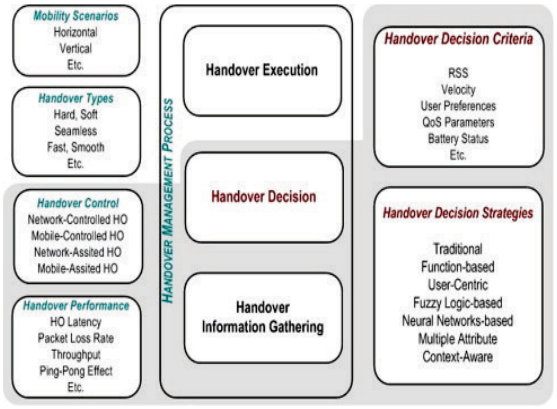


Figure 1: Handover concept [19]

B. Modulation Strategies

Modulation strategies are used in wireless data transmissions to convert digital data into analog signals; these signals are then transmitted over a carrier frequency and restored back to its digital state upon arriving at the receiver. It does this by defining the data rate per second and Forward Error Correction (FEC) used in transmission bursts [20]. Bits of data are encoded into symbols which are unique combinations of phase, slight frequency variations and amplitude. The symbols are encoded with bits of data. Higher modulation strategies can achieve a larger number of bits per symbol. However, with higher bit-rate strategies it is more difficult, particularly in high noise conditions such as bad weather, to distinguish between the large amounts of different possible symbols. For example, low Signal-Noise Ratio (SNR) connections can lead to a larger bit-error ratio (BER) when the signal received is too low for the receiver to decode it accurately.

To mitigate these problems and ensure an adequate QoS in the link, dynamic selection algorithms are implemented inside wireless devices. When the quality of the wireless connection deteriorates, the modulation scheme used is changed to a lower bit rate scheme such as QPSK and BPSK. The advantage of lower bit rate modulation schemes is a lower BER but this comes at the cost of a lower data rate and, thus, throughput. In ideal conditions, higher data-rate strategies can be selected but as signal strength is reduced due to increased distance and/or interfering phenomena, dynamic algorithms present in most wireless devices can select a lower data-rate strategy to optimize QoS in the connection. Figure 2 illustrates the BER for various modulation strategies level as a function of short-term average SNR.

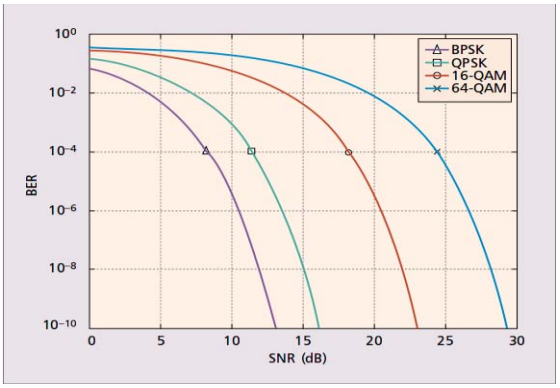


Figure 2: BER/SNR correlations [21]

The selection of modulation strategy is dependent on several metrics both performances related and external. These include distance from the access point, available bandwidth and the amount of noise present on the network. Interference due to weather

conditions such as rain fade can cause "scattering" to occur. In this instance, the signal is scattered in all directions, thus reducing its overall strength and make it more difficult to decode a state when it reaches the receiver [22]. This behavior is exaggerated on higher modulation strategies because the larger numbers of symbols can be more difficult to differentiate at the receiver, particularly in noisy conditions, leading to a higher BER. For this reason, methods are continuously being developed to ensure the proper modulation strategy selection to ensure QoS [23].

C. Artificial Neural Networks

Synaptic Weights that are used for our weight configurations are special variables that are part of an Artificial Neural Network (ANN). These weights can be compared to the biological systems that make up a human brain in terms of how one learns to achieve a goal or solve a problem. A collection of neurons work together to solve a problem and will learn while attempting to solve it. I.e. After a "cycle" of learning, more "weight" is added to a certain neuron to indicate its higher importance. The main features of an ANN are its adaptive learning capabilities, flexibility in dealing with new environments as well as robustness and fault tolerance. Neural networks can only learn through repetition and cannot be "programmed" to know the answer or optimal result. These adjustable synaptic weights are applied to each input in the neural network and the sum of these weighted inputs is then compared to a threshold value. Exceeding this value fires a learning algorithm which applies corrections to the weights. These adjusted weights are then applied in the next cycle. The neural network works in such a way that if a particular input is found to have a positive effect on the output of the network, than the weight upon that value is increased to reflect its high importance.

For this paper, there will be no training of a neural network; rather previous results from brute force testing of every weight configuration from 0 to 1 in increment of 0.1 shall be used. We shall take a selection of these results from these brute force tests and use them for the tests for video streaming.

D. Evalvid

Evalvid is a comprehensive video quality evaluation tool-set [24]. It analyses the video quality metrics of media transmitted over wireless networks such as video frames lost over transmission, Peak Signal Noise Ratio (PSNR) and a video quality evaluation metric. It currently supports a variety of video codecs including H.264, MPEG-4 and H.263. The usage of Evalvid to evaluate video involves calculating the PNSR of a video before it is streamed out over the network. This gives us a reference PSNR file that shows the performance of the original video file and, as such, the output file can be compared to the original source file in order to measure network performance. The file is then streamed across a network, using either methods such as TCPdump, NS-2 or NS-3. These methods will leave dump files for the sender and receiver, the dump files are then used to reconstruct the original video including any errors and loss frames.

From this video, the file is converted back into a raw video format (either a YUV or Y4M file), PSNR values are calculated and the Evalvid Mean Opinion Score (MOS) function is used to objectively state the overall quality of the received video.

IV. EXPERIMENT

A. Setup

A series of simulations are undertaken in which the QPSK modulation strategy is applied to all transmitters within a heterogeneous network configuration consisting of three WiMax networks and a back-up ubiquitous LTE network. The simulations are run using the NS-3 network simulator and its built-in WiMax module [26]. The topology used for this simulation had three transmitting WiMax base-stations positioned at intervals along a straight line route. The WiMax base stations are located ten meters off road, at a height of thirty meters and broadcasted at a power level of approximately +43dBm while the user's mobile device broadcasted at +23dBm. The LTE network has a set delay and error rate while the WiMax network had a varying delay and error rate. The route takes 600 seconds to traverse while a 500 MB size MP4 video file is streamed to the user's mobile device.

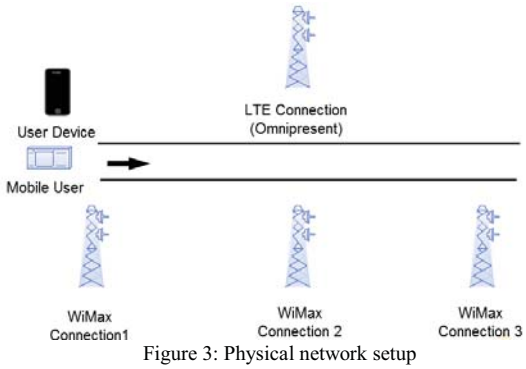


Figure 3: Physical network setup

On a second-by-second basis, the handover algorithm takes in the performance metrics (loss, delay, RSS, jitter and SNR) of the three WiMax networks, normalizes them, and applies configurable synaptic weights onto them (normalized input * weight value). These values are then added together to form a sum.

The WiMax link with the highest sum that has also exceeded the threshold value of the algorithm is selected at that second and the stream is handed off to it. If the threshold fails to be exceeded by any of the WiMax links, then the network falls back to the back-up LTE connection. At the conclusion of the simulation the received video file was compared to the original file so that a mean opinion score (MOS) can be calculated. Assuming poor weather conditions, the above mentioned modulation strategy supported a data rate of 13.82 Mb/s. The Mean Opinion Score value for the entire received video is ranked as such:

Table I: MOS Quality Chart

MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very Annoying

The sixteen weight configurations used in these experiments are taken from previous brute force testing done in the same topology, using all possible weight configurations between 0 and 1 that are to the first decimal place. All sixteen configurations come from previous brute force testing done on the network. For these tests, UDP packet streaming is used instead of a multimedia stream, with the final throughput used as an indicator of overall performance. 815 tests were ran and then ranked according to their final throughput. From the results as seen in figure 4, the results can be divided into three areas of performance. These three sections are the low performance section where the weights are too low to cause handovers to any of the WiMax connections, the degrading performance section in which handovers to WiMax begin to occur more frequently and the high performance section where the WiMax connections are used as the access points for the entirety of the simulation.

At the start of the simulation the connection used will be either a WiMax link that has surpassed the threshold value at the start or the back-up LTE link used for the simulation. If it occurs at any second that the normalized performance metric values coupled with their respective weight configuration fails to stimulate the neurons of the handover algorithm enough to exceed the threshold value, the user device will handover back to the back-up LTE link. These sixteen tests will be run four times, with the quality of the LTE link degrading further each time.

Table II: LTE link setups used for Simulations

LTE Quality	Error Rate	Delay
Best	4%	5ms
Average	8%	15ms
Bad	10%	30ms
Worst	30%	200ms

The weight configurations used for these tests are based off previous work where a brute force method was applied to see which weight configuration within a range of between 0 and 1 will result in the best throughput in a similar topology to the one used for these experiments. Three performance areas representing the increase of throughput are found from the 815 weight configurations that are used as seen in Figure 5. For these tests, a selection of weight configurations from each area of performance is chosen and used for these experiments. In the case of test 0, the weights being set to zero will mean the back-up connection is used at all times.

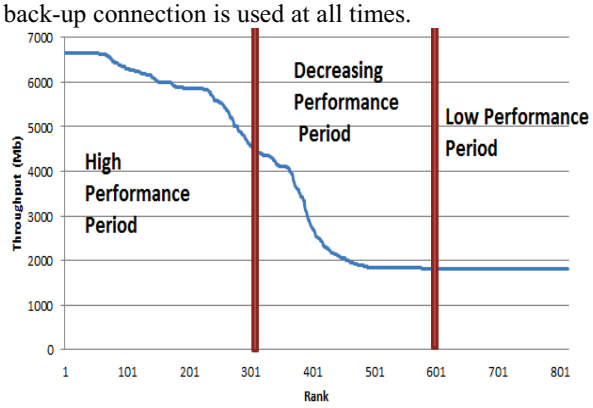


Figure 4: Performance periods from previous brute force testing

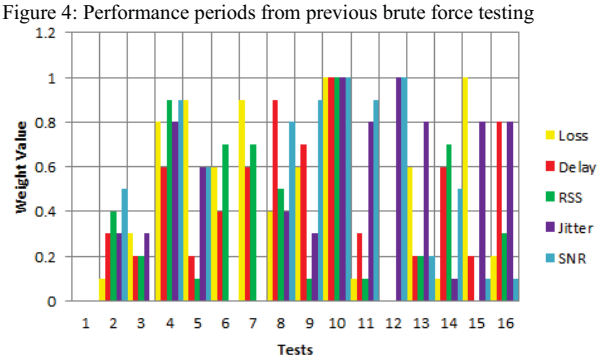


Figure 5: Weight configurations for each test

Table III: Weight Configurations used for all tests

Test	Loss	Delay	RSS	Jitter	SNR
1	0	0	0	0	0
2	0.1	0.3	0.4	0.3	0.5
3	0.3	0.2	0.2	0.3	0
4	0.8	0.6	0.9	0.8	0.9
5	0.9	0.2	0.1	0.6	0.6
6	0.6	0.4	0.7	0	0
7	0.9	0.6	0.7	0	0
8	0.4	0.9	0.5	0.4	0.8
9	0.6	0.7	0.1	0.3	0.9
10	1	1	1	1	1
11	0.1	0.3	0.1	0.8	0.9
12	0	0	0	1	1
13	0.6	0.2	0.2	0.8	0.2
14	0.1	0.6	0.7	0.1	0.5
15	1	0.2	0	0.8	0.1
16	0.2	0.8	0.3	0.8	0.1

B. Results

Table IV: Connection handovers per test.

Test	WiMax 1	WiMax 2	WiMax 3	LTE
1	0	0	0	1
2	1	0	0	1
3	0	0	0	1
4	1	0	1	1
5	1	8	8	1
6	13	18	6	2
7	19	27	9	1
8	1	0	1	1
9	1	8	8	1
10	1	0	1	1
11	1	0	0	1
12	1	0	0	2
13	1	0	0	1
14	3	1	1	1
15	1	8	8	1
16	1	0	0	1

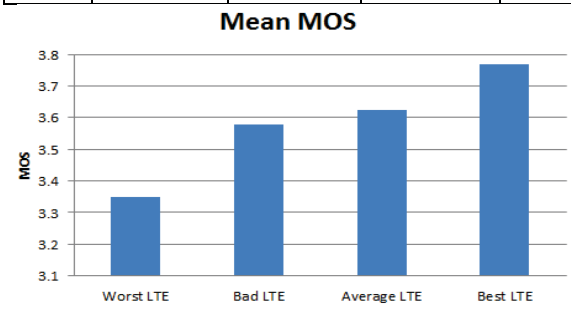


Figure 6: Mean MOS values for tests on all LTE simulations

Figure 6 illustrates the mean MOS values of test results across all 4 LTE simulations. We can see that the worst the LTE, the worse the lower the average MOS. Figure 7 illustrates the differences between all four connections qualities used for the back-up LTE network in regards to how it can affect the final MOS result for a specific test.

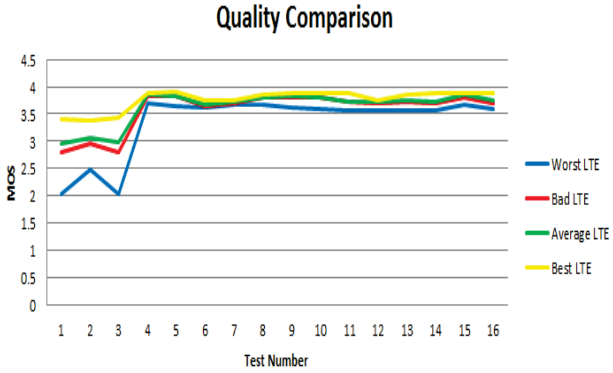


Figure 7: A comparison of all tests across all LTE simulations

It can be seen that the best quality LTE simulation outputs better quality video than the other LTE simulations even when the back-up link is relied on as seen in Test 1-4. What is evident across all four LTE quality scenarios is that handing off to a WiMax connection near or at the start of the simulation guarantees a better final MOS score for the video output than using the LTE connection exclusively. This is due to the normalized performance metrics multiplied by their weight values almost always being enough to prevent fallback to the back-up LTE connection at any point of the simulation.

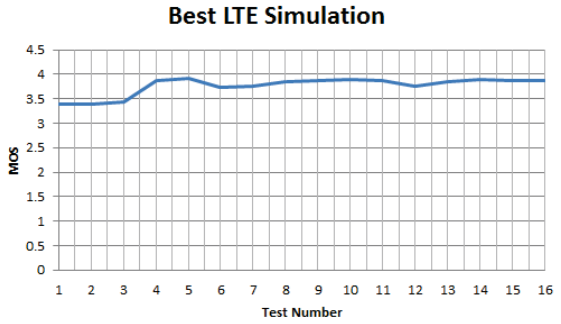


Figure 8: MOS results for best LTE simulation

In the best quality LTE simulation, the weight configurations used did not have a major effect in the final MOS value in comparison to the MOS result of the pure LTE test. This can be explained by the LTE connection being similar in quality of service to the WiMax connections. As such, it is difficult to tell what weighting configuration is responsible for best performance in this case as most tests can have a high

MOS value as long as the weights are sufficiently stimulated enough for handovers to occur. The fact that a pure LTE result results in a MOS value of 3.4 indicates there is not a big gap between that and the results generated through handover applications. In this case, all that a higher MOS result requires is moderate to high weightings to make the WiMax alternatives capable of exceeding the algorithms threshold.

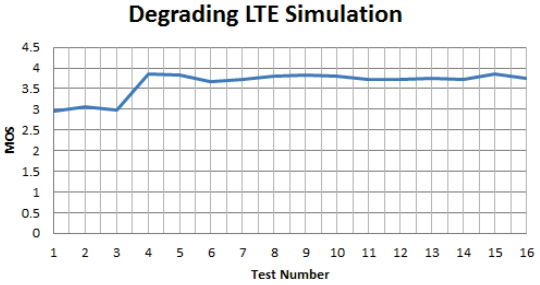


Figure 9: MOS results for Average LTE simulation

For the degrading simulation, the application of the handover algorithm resulted in a noticeable increase in the potential MOS result. The pure LTE test returned a MOS of 2.96, meaning the video will have noticeable errors present. If a high weight value is applied to loss and/or jitter, the final MOS value can increase by over half a rank. The MOS value becomes stable at around 3.8 in tests 11-16, where high value weights are applied to at least two of jitter, delay or SNR. This concurs with the findings of [9] and [15] in regards to the importance of jitter and delay as performance metrics for video streaming.

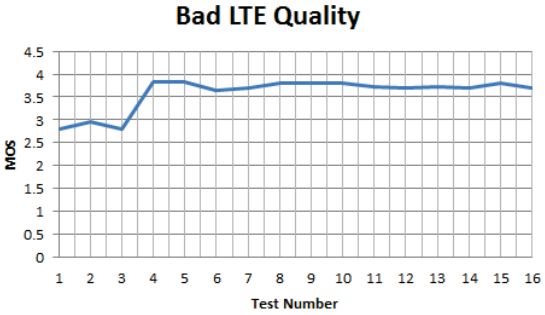


Figure 10: MOS performance for the bad LTE simulation

For the bad LTE simulation, the pure LTE scenario results in the MOS dipping to 2.79 in the case of no handovers occurring. This means the outputted video has noticeable errors present. Applying the weights correctly can raise the MOS up to around to near rank 4, meaning the handover algorithm can potentially improve streamed video quality by a whole rank. As in the average LTE simulation, the best results are due to the application of moderate to high weight values onto jitter, delay or SNR.

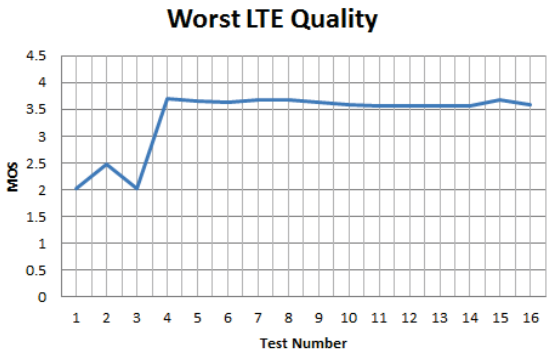


Figure 11: MOS performance for worst LTE simulation

For the worst quality scenario, the MOS has fallen to rank 2 on the pure LTE test, meaning without handovers the video quality is quite poor and the errors are both noticeable and quite annoying. Applying the weights to favor jitter, delay and SNR in particular can raise up to over 3.5, improving quality to the point the video quality is fair and raising the video's overall rank by a large 1.5 value.

In summary, the handover algorithm is best used when the weight values are applied with a large focus on delay, jitter and SNR. These are the most important metrics for quality video streaming. However, at the same time it is also vital to apply moderate to high value weightings to these metrics as well, this allows the algorithm to not only identify the best connection but also to inform the handover algorithm that this connection is ideal for streaming video data and reduce the amount of times it has to fall back to the LTE connection for the video stream.

V. CONCLUSION

The purpose of this paper is to illustrate how the implementation of a performance metrics based handover algorithm can have a large improvement on the video quality of video files that are streamed through a wireless network topology. We defined four network topologies with the only difference being the quality of the LTE connection that is used as a backup for the three WiMax networks. If the performance metrics of any of the three alternative WiMax networks coupled with a unique weight configuration exceed the defined threshold value of the performance algorithm, a handoff is initiated to the best connection. The results of these tests illustrate that relying exclusively on a single link without properly considering the possibility of superior alternative connections for video streaming on the network can result in a poorer quality output at the receiver. In conclusion, for video streaming over heterogeneous networks, it is important to implement a handover algorithm that also considers the important performance metrics for video streaming,

namely jitter and delay, to ensure that the best quality connection is used when video streaming. In the case of bad quality connection unsuitable for quality video streaming, a handover algorithm can potentially raise the mean opinion score of a received video file to a higher rank as opposed to using a single degraded connection. These experiments have also demonstrated the importance of Jitter, delay and SNR as input metrics into the decision algorithm for what network to connect to in a point in time.

REFERENCES:

[1] Claypool, Mark, and Jonathan Tanner. "The Effects of Jitter on the Perceptual Quality of Video." In Proceedings of the Seventh ACM International Conference on Multimedia (Part 2), 115–118. MULTIMEDIA '99. New York, NY, USA: ACM, 1999. doi:10.1145/319878.319909.

[2] Angrisani, L., K. Kyriakopoulos, A. Napolitano, D.J. Parish, M. Vadursi, and W. Whittow. "An Experimental Analysis of the Effects of Noise on Wi-Fi Video Streaming." In Instrumentation and Measurement Technology Conference (I2MTC), 2010 IEEE, 1551–1555, 2010. doi:10.1109/IMTC.2010.5488062.

[3] Kalman, M., E. Steinbach, and B. Girod. "Adaptive Media Playout for Low-Delay Video Streaming over Error-Prone Channels." Circuits and Systems for Video Technology, IEEE Transactions on 14, no. 6 (June 2004): 841–851. doi:10.1109/TCSVT.2004.828335.

[4] Al-Majeed, S.S., M. Fleury, and S. Janardhanan. "Effective Broadband Video Streaming during Wireless Vertical Handovers." In Consumer Electronics (ICCE), 2012 IEEE International Conference on, 357–358, 2012. doi:10.1109/ICCE.2012.6161900.

[5] Institute of Electrical and Electronics Engineers, "IEEE Standard for Local and metropolitan area networks – Part 21: Media Independent Handover Services", LAN/MAN Standards Committee of the IEEE Computer Society, 21 Jan. 2009

[6] Larsson, A.; Piotrowski, A.; Giles, T.; Smart, D., "Near-earth RF propagation - Path loss and variation with weather," Radar (Radar), 2013 International Conference

[7] Institute of Electrical and Electronics Engineers, "IEEE Standard for Local and metropolitan area networks – Part 21: Media Independent Handover Services", LAN/MAN Standards Committee of the IEEE Computer Society, 21 Jan. 2009.

[8] Kesavan, U.; Tharek, A. R.; Rahim, S. K A; Rafiqul, I.M., "Propagation studies on rain for 5.8 GHz and 23 GHz point to point terrestrial link," Computer and Communication Engineering (ICCCE), 2012,

[9] Claypool, Mark, and Jonathan Tanner. "The Effects of Jitter on the Perceptual Quality of Video." In Proceedings of the Seventh ACM International Conference on Multimedia (Part 2), 115–118. MULTIMEDIA '99. New York, NY, USA: ACM, 1999.

[10] Yokota, M.Y.S., and T. Nakajima. "A Quality Evaluation of High-Speed Video Streaming on Congested IP Networks." In Information and Telecommunication Technologies, 2005. APSITT 2005 Proceedings. 6th Asia-Pacific Symposium on, 53–58, 2005. doi:10.1109/APSITT.2005.203630.

[11] Trestian, R., A.-N. Moldovan, O. Ormond, and G. Muntean. "Energy Consumption Analysis of Video Streaming to Android Mobile Devices." In Network Operations and Management Symposium (NOMS), 2012 IEEE, 444–452, 2012. doi:10.1109/NOMS.2012.6211929.

[12] Chang, Shih-Ying, Hsin-Ta Chiao, Xin-Yan Yeh, and Ming-Chien Tseng. "UDP-Based File Delivery Mechanism for Video Streaming to High-Speed Trains." In Personal Indoor and Mobile Radio Communications (PIMRC), 2013 IEEE 24th International Symposium on, 3568–3572, 2013. doi:10.1109/PIMRC.2013.6666768.

[13] Hatakeyama, K., S. Tsumura, and S.-i. Kuribayashi. "Packet Transmission Control of Preventing the Perceptual Video Quality Deterioration in All IP-Based Network." In Information Networking, 2008. ICOIN 2008. International Conference on, 1–4, 2008. doi:10.1109/ICOIN.2008.4472816.

[14] Papadimitriou, P., and V. Tsaoussidis. "QRP04-4: End-to-End Congestion Management for Real-Time Streaming Video over the Internet." In Global Telecommunications Conference, 2006. GLOBECOM '06. IEEE, 1–5, 2006. doi:10.1109/GLOCOM.2006.438.

[15] Cunningham, G., S. Murphy, L. Murphy, and P. Perry. "Seamless Handover of Streamed Video over UDP between Wireless LANs." In Consumer Communications and Networking Conference, 2005. CCNC. 2005 Second IEEE, 284–289, 2005. doi:10.1109/CCNC.2005.1405184.

[16] Cunningham, G., P. Perry, J. Murphy, and L. Murphy. "Seamless Handover of IPTV Streams in a Wireless LAN Network." Broadcasting, IEEE Transactions on 55, no. 4 (December 2009): 796–801. doi:10.1109/TBC.2009.2030466.

[17] Ramli, K.; Wicaksono, A.; Budiardjo, B.; Sari, R.F., "QoS experimentation analysis on the impact of user mobility on video streaming application over mobile IPv6 network,"Networks, 2005. Jointly held with the 2005 IEEE 7th Malaysia International Conference on Communication., 2005 13th IEEE International Conference on , vol.1, no., pp.6 pp.,, 16-18 Nov. 2005

[18] Stein J, "Survey of IEEE802.21 Media Independent Handover Services" Available to download at: http://www.cs.wustl.edu/~jain/cse574-06/ftp/handover/

[19] Meriem Kassar, Brigitte Kervella, Guy Pujolle, An overview of vertical handover decision strategies in heterogeneous wireless networks, Computer Communications, Volume 31, Issue 10, 25 June 2008, Pages 2607-2620, ISSN 0140-3664, http://dx.doi.org/10.1016/j.comcom.2008.01.044.

[20] Marabissi, D.; Tarchi, D.; Fantacci, R.; Genovese, F., "Adaptive Modulation Algorithms based on Finite State Modeling in Wireless OFDMA Systems,"Personal, Indoor and Mobile Radio Communications, 2007. PIMRC 2007.

[21] Catreux, S.; Erceg, V.; Gesbert, D.; Heath, R.W., "Adaptive modulation and MIMO coding for broadband wireless data networks,"Communications Magazine, IEEE , vol.40, no.6, pp.108,115, Jun 2002

[22] Harb, K.; Srinivasan, A.; Changcheng Huang; Cheng, B., "Prediction method to maintain QoS in weather impacted wireless and satellite networks,"Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on

[23] Harb, K.; Changcheng Huang; Srinivasan, A.; Cheng, B., "Intelligent Weather Aware Scheme for Satellite Systems," Communications, 2008. ICC '08. IEEE International Conference on

[24] "Evalvid - A Video Quality Evaluation Tool-set" Availiable at: http://www2.tkn.tu-berlin.de/research/evalvid/

[25] "How to evaluate MPEG video transmission using the NS2 simulator?" Available to download at: http://hpds.ee.ncku.edu.tw/~smallko/ns2/Evalvid_in_NS2.htm

[26] Jahanzeb Farooq and Thierry Turletti. 2009. An IEEE 802.16 WiMAX module for the ns-3 simulator. 2nd International Conference on Simulation Tools and Techniques (Simutools '09)

PLAAO – Perceptron Based Load Balancing Algorithm Using Antenna Orientation

Mikel Zuzuarregui Ibarbia [1,2], Enda Fallon [1], Yuansong Qiao [1], Paul Jacob[1], Sajeevan Achuthan[2]

[1] Software Research Institute, Athlone Institute of Technology, Ireland
[2] Ericsson Research, Athlone, Ireland
mzuzuarregui@research.ait.ie, efallon@ait.ie, ysqiao@research.ait.ie, pjacob@ait.ie, sajeevan.achuthan@ericsson.com

Abstract

For network operators, balancing the throughput between users is important to accomplish a successful Quality-of-Service (QoS). There are number of parameters which can be altered to achieve overall acceptable QoS. In this paper, we consider changing the antenna orientation to maximize throughput. Our approach captures and optimises antenna orientation based on historic performance using a perceptron Artificial Neural Network (ANN) approach. The approach considers a heterogeneous network scenario with multiple cell propagation from a radio base station. Using system simulations (NS-3) of a sample LTE, we verify that the system is learning. PLAAO is writing in R to become pluggable with other systems and it is extensible. Results illustrate improved collective throughput utilising the PLAAO algorithm.

Keywords: Long Term Evolution (LTE), Load Balancing, Directed Learning

1 Introduction

Over the last years, the complexity of cellular network architectures has increased. The increased complexity is driven by significant consumer interest in mobile devices. The number of mobile data subscribers is expected to grow to 5 billion by 2016 [1].

In order to address the increased complexity of mobile networks a number of research areas have emerged. Self Organizing Networks (SONs) is one such area. SON involves the optimisation of planning configuration, management, optimization and healing of the network. SON was first introduced in LTE in 2011 [2].

In this paper we propose PLAAO – A Perceptron Based Load Balancing Algorithm Using Antenna Orientation. PLAAO is a SON approach which uses historic network behaviour in order to optimise network performance. In particular PLAAO alters antenna orientation in order to optimise collective throughput for all Mobile Nodes (MNs) in a cell. Having outlined the structure of the PLAAO algorithm this paper investigates; the learning efficiency of the approach, the optimisation of throughput achievable with the approach and the interrelation of loss, delay, throughput and antenna orientation?

This paper is organised as follows. Section 2 outlines related work in the area. Section 3 discusses the algorithm of the learning mechanism. The experimental setup and simulation result are explained in section 4. Finally, in section 5 the conclusions are presented.

2 Related Work

2.1 Self-Organizing Network (SON)

As radio networks such as LTE are becoming more complex the necessity for some degree of self-organisation is increasing. Traditional network management approaches had a high level of human intervention. As the complexity of networks increase such levels of human interaction are not practical. Self-Organising Networks (SON) enable networks to self-configure, self-optimize and self-heal. Using SON telecommunications networks can implement self-optimization. From the business logic viewpoint, benefits are realised in the area of capital expenditure (CAPEX) and later operational expenditure (OPEX).

The most relevant research in the area of SON is concerned with handover optimization and management [4][5][6], interference management [7][8][9], coverage/capacity optimization, load management [10][11][12] and self-configuration and planning [13].

In [4] the authors proposed a modified Self-Organizing Map (SOM) mechanism which is used to allow a femtocell to learn the locations of the indoor environment from where handover requests have occurred. Based on previous experience the approach considers whether to permit or prohibit these handovers. In [5] different parameter optimization levels (network-wide, cell-wide, and cell-pair-wide) and the impact of measurement errors have been considered. The proposals utilise SON approaches for handover optimization. In [6] the authors proposed a novel handover optimization mechanism. To achieve that, each cell in a RAT updates its handover parameters in an autonomous and automated manner depending on its traffic and mobility conditions. The algorithm uses a feedback controller to update the handover parameters as a means to providing a steady improvement in the network performance.

In [7] the authors propose a SON viewpoint by optimizing fractional frequency reuse (FFR) and adapting to dynamic traffic maps. Their approach addresses the problem as an optimization with multiple key performance indicators (Multi-KPIs), and a traffic-based dynamic spectrum management (DSM) algorithm. On the other hand, in [8] the authors propose a completely self-organizing approach for the small-cell base station (SBS). They use a novel cognitive interference alignment-based scheme to protect the macrocell from cross-tier interference while mitigating the co-tier interference. In [9] they assess the viability of heterogeneous networks composed of legacy macrocells, which are underlaid with self-organizing picocells. For that purpose, a comprehensive analytical framework is introduced to evaluate the performance of such self-organizing networks in terms of outage probability and average channel capacity with respect to the tagged receiver.

In [11] the authors proposed a technique to compute desirable user associations by an interference model that explicitly takes base-station loads into account. Based on the ability to predict cell loads, they derive algorithms that jointly adapt user-association policies and antenna-tilt settings for multiple cells. In [12] the authors present a traffic-light-related approach to autonomous self-optimization of tradeoff performance indicators in LTE multitier networks. Introducing a low-complexity interference approximation model, the related optimization problem is formulated as a mixed-integer linear program and is embedded into a self-organized network operation and optimization framework.

For the self-configuration and planning in [13] the authors present an automatic method for defining when and how to change an existing tracking areas TA plan to minimize network signalling. In this paper the TA replanning problem is formulated as a classical graph partitioning problem, which is then solved by a sophisticated graph partitioning algorithm combining multilevel and evolutionary techniques.

2.2 Rule-based Systems

Proposed in the 1960's, rule-based systems gained popularity in the 1970's and 1980's [14]. Rule based systems are intended to replicate human expertise. The systems are based on rules and are used in repeatable scenarios to make intelligent and quick decisions.

A rule system has five basic components: Data-base Matching, Condition-Action Rule, Rewriting Rules, Forward / Backward Chaining, Arbitration. There are able to offer relations (if-then), recommendations (if-and-and-and-then), directives (if-and-then), strategies (if-then-if-and-then) and heuristics (if-and-and-then). [10] presents a rule-based system connected with a neural network. The author proposed a rule-based reinforcement learning approach with the aim of reducing congestion issues.

2.3 Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) use mathematical formulae to replicate nervous system operations by learning patterns and relationships in data. The learning rules enable the network to gain knowledge from available data and apply that knowledge to assist in decision making. Such learning depends heavily on relating previous events to predict future situations. The ANN can consist of many nodes similar to neurons in the brain. Each node has an associated function which, along with a set of input parameters, determines the output of the node. Modifying the input parameters may alter the node function. The output signal is obtained by applying activations to the network inputs.

Neural network approaches are generally defined in 2 classes; supervised and unsupervised. Unsupervised approaches are motivated by the requirement to be autonomous self-organizing structures. Such an approach can be generally considered as the clustering of input data in order to extract useful information. Supervised neural networks are generally motivated by the requirements of a specific task. As the problem domain is well defined supervised approaches tend to optimize specific performance criteria.

This paper proposes a direct learning algorithm to automatize the antenna orientation of the radio base station. The approach uses feed-forward network architecture with a back-propagation algorithm. The aim of the system is to maximise overall throughput by probing network characteristics.

There are many studies around artificial neural network, especially in learning based network load balancing for LTE networks [15] [16]. In [15] it is proposed a dynamic ANN schema to achieve load balancing. In [16] the author propose a modified self-organising map (SOM). They used it to reduce the total number of handovers.

2.3.1 Feed-Forward Neural Networks

The first work on ANN was presented by Mc Cullock and Pitts in 1943 [17]. They recognized that combining many simple processing units together could lead to an overall increase in computational power. The basic idea of a McCulloch-Pitts model is to use components which have some of the characteristics of biological neurons. A biological neuron has a number of inputs which are "excitatory" and some which are "inhibitory". What the neuron does depends on the sum of inputs. The excitatory inputs tend to make the cell fire and the inhibitory inputs stop it firing. The work proposed a Threshold Logic Unit (TLU) which used weighted binary inputs.

The McCulloch and Pitts network had a fixed set of weights and it was Hebb [18] who developed the first learning rule. His premise was that if two neurons were active at the same time then the strength between them should be increased. Hebbian learning involves weights between learning nodes being adjusted so that each weight better represents the relationship between the nodes. The weight between two neurons increases if the two neurons activate simultaneously. The weight between two neurons reduces if they activate separately. Nodes that tend to be either both positive or both negative at the

same time have strong positive weights, while those that tend to be opposite have strong negative weights.

The following formula describes Hebbian learning:

$$w_{ij} = \frac{1}{p}\sum_{k=1}^{p} x_i^k x_j^k \quad (1)$$

w_{ij} is the weight of the connection from neuron j to neuron i, p is the number of training patterns, and x_i^k is the k^{th} input for neuron i.

2.3.2 Directed Learning

A directed learning algorithm shares some of the characteristics of both supervised and unsupervised ANN approaches. Weight alteration can be implemented in a manner similar to Hebbian learning. However the algorithm is unsupervised as it does not implement a specific training phase. At the beginning the system try to get the highest throughput with any orientation and following some rules suggest the next orientation. Depending on attained throughput, the weights are adjusted. PLAAO is a directed learning approach. Some studies evaluate the effect of antenna tilt on base station performance. In [19] the authors propose reinforcement learning based on unsupervised machine learning. The system can learn from the system parameter changes.

3. PLAAO – Perceptron Based Load Balancing Algorithm Using Antenna Orientation

This paper proposes a Perceptron Based Load Balancing Algorithm Using Antenna Orientation algorithm (PLAAO). PLAAO captures and optimises antenna orientation based on historic performance using a perceptron ANN approach. The approach considers a heterogeneous network scenario with multiple cell propagation from a radio base station.

The PLAAO approach consists of X_i neuron inputs corresponding to the performance metrics delay and loss. Each of the values for delay and loss were normalized on a scale of 0 (Worst performance) to 100 (best performance). Each weighted performance metric is multiplied by a corresponding synaptic weight W_i. The product of neuron input and synaptic weight is the activation value, V_y. Each neuron compares the activation value to a linear threshold. If the activation value exceeds the threshold a particular antenna orientation is selected.

$$V_y = \sum_{i=1}^{N} X_i \cdot V_i \quad (2)$$

The following equation outlines the relationship between antenna orientation in degrees of rotation and activation value. The classification of activation values is mutually exclusive in each neuron. For example Neuron 1 considers an activation value in the range 0 to 0.1. Neuron 1 is the only neuron which will result in an antenna rotation of -80°.

$$\varphi_y = \begin{cases} -90\ if\ V_i \le 0 \\ -80\ if\ 0 < V_i \le 0.1 \\ -70\ if\ 0.1 < V_i \le 0.2 \\ -60\ if\ 0.2 < V_i \le 0.3 \\ -50\ if\ 0.3 < V_i \le 0.4 \\ -40\ if\ 0.4 < V_i \le 0.5 \\ -30\ if\ 0.5 < V_i \le 0.6 \\ -20\ if\ 0.6 < V_i \le 0.7 \\ -10\ if\ 0.7 < V_i \le 0.8 \end{cases} \qquad \varphi_y = \begin{cases} 0\ if\ 0.8 < V_i \le 0.9 \\ 10\ if\ 0.9 < V_i \le 1 \\ 20\ if\ 1 < V_i \le 1.1 \\ 30\ if\ 1.1 < V_i \le 1.2 \\ 40\ if\ 1.2 < V_i \le 1.3 \\ 50\ if\ 1.3 < V_i \le 1.4 \\ 60\ if\ 1.4 < V_i \le 1.5 \\ 70\ if\ 1.5 < V_i \le 1.6 \\ 80\ if\ 1.6 < V_i \le 1.7 \\ 90\ if\ V_i > 1.7 \end{cases} \quad (3)$$

The PLAAO algorithm is intended to optimize collective throughput for all Mobile nodes (MNs) in a base station cell. As PLAAO is an unsupervised approach, changes in antenna must be analysed over time to determine best performance. There is no target throughput against which PLAAO can be trained. The rate of change, c, of throughput is calculated. $c = \frac{\sum(x-x')(y-y')}{\sum(x-x')^2}$ (4)

Based on c the system can change the weights for the next iteration and proceed with learning. A positive c means improved throughput in comparison with the previous cycle. A negative c means that the value of throughput is lower than in the previous iteration.

In order to control the rate of learning PLAAO defines a user configurable learning rate constant r. This constant takes values from 0 to 1. This parameter makes softer the learning for the system. It arrives to the optimal value slower. We define the error correction, ΔW, as the product of c and r.

4 Simulated Evaluation of the PLAAO Approach

This section analyses the network performance of PLAAO approach in order to optimize collective throughput for all MNs in a base station cell. The simulation is undertaken using the NS-3 open source platform using LTE libraries. The R language is used to efficiently construct ANN learning tables.

The following parameters were configured in the simulation; Power : 20dbm, Noise: 0, Number of base station = 1, Number of nodes: 7, Data rate: 2Mb/s, Packet size: 900 bytes, Antenna model: cosine, Beam-width: 60 degrees and Max Gain: 0

The Figure 1 illustrates how the base station and nodes are organized along the x and y axis.

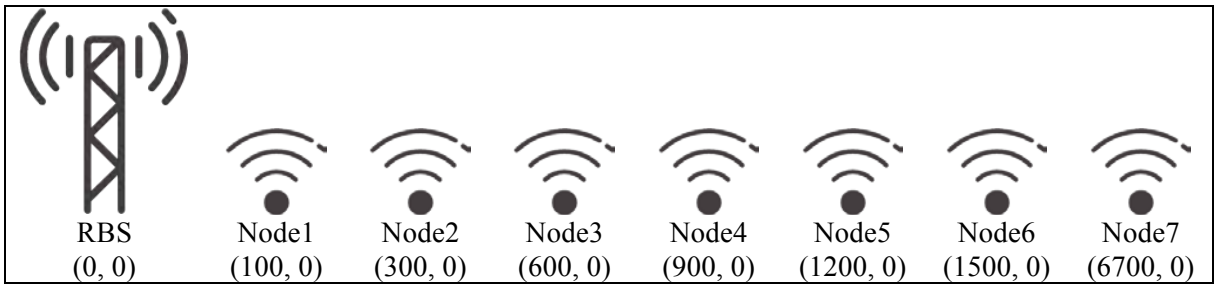


Figure 1: Network Topology - Positions are in brackets (x, y).

Depending on the orientation of the antenna, some nodes are in or out of the range. While other parameters can affect antenna range; transmission power, noise etc, this work focused on antenna orientation.

In order to control the rate of learning PLAAO defines a user configurable learning rate constant r. The selection of an appropriate learning rate is critical for the effective operation of the algorithm. If the learning rate is too low the network learns very slowly. If the learning rate is too high weights diverge, resulting in little learning. We define the error correction, ΔW, as the product of c and r.

Table 1 illustrates the recorded throughput for each of the 19 antenna orientation configurations. The scenario is symmetric so the orientation, so as an example, 40^o and -40^o have the same throughput.

Orientation	0	10	20	30	40	50	60	70	80	90
Throughput	9.561	9.561	9.053	8.384	7.453	5.993	4.595	2.896	1.327	0.573

Table 1: Throughput for Each Antenna Orientation

The best throughput is achieved with the orientations -10^o and 10^o. The throughput for the orientations 20^o and -20^o are also acceptable.

Table 2 illustrates the activation value, $V_{y,}$ achieved for the 11 training iterations of the PLAAO algorithm. Input values are normalized between 0 (worst performance) to 1 (best performance). The first learning iteration results in an activation value of V_y = (.851 · 0.174) + (.519 · .092) = 0.195856879. As this activation value falls in the range 0.1 to 0.2 an antenna orientation of -70^o is selected for the next learning cycle. The second learning iteration results in an activation value of 0.482609417. As this activation value falls in the range 0.4 to 0.5 an antenna orientation of -40^o is selected for the third learning cycle. The activation value is reaching the value 1 iteration by iteration (this is the optimal value range .7 – 1). On the other hand, as it is mentioned before, the normalized delay and loss are near from the optimal values.

		Inputs		Normalized		Weights		Output
Iteration	**Orientation**	**Delay**	**Loss per cent**	**Normalize Delay**	**Normalize Loss**	**W1**	**W2**	**Activation Value ($V_{y)}$**
1	40	10.443798	48.11441064	0.850802886	0.518855894	0.17392098	0.092288451	0.195856879
2	-70	20.326	79.6259185	0.709628571	0.203740815	0.54659298	0.464960451	0.482609417
3	-40	10.443798	48.11441064	0.850802886	0.518855894	0.31870898	0.237076451	0.394167034
4	-50	10.698247	58.18303273	0.8471679	0.418169673	0.54659298	0.464960451	0.657488387
5	-20	0.920868	37.01949353	0.986844743	0.629805065	0.47356448	0.391931951	0.714175345
6	-10	0.700923	33.50337038	0.989986814	0.664966296	0.62656698	0.544934451	0.982656092
7	10	0.700923	33.50337038	0.989986814	0.664966296	0.65195298	0.570320451	1.024668732
8	20	0.920868	37.01949353	0.986844743	0.629805065	0.65195298	0.570320451	1.00256708
9	20	0.920868	37.01949353	0.986844743	0.629805065	0.62656698	0.544934451	0.961526808
10	10	0.700923	33.50337038	0.989986814	0.664966296	0.62656698	0.544934451	0.982656092
11	10	0.700923	33.50337038	0.989986814	0.664966296	0.65195298	0.570320451	1.024668732

Table 2: Initial performance metrics, normalized metrics, weights and output

Table 3 illustrates the learning implemented by the PLAAO algorithm when the learning rate r =.05. The initial weights w1=.174 and w2=.0922 and orientation=40 are randomly allocated. Table 2 and Table 3 outlines that this initial weight configuration are far from optimal. As a consequence, there is a quite a long training with 11 cycles. To calculate the rate of change, c, of a linear regression line through two points, we assume that the initial throughput was 0. By calculating $c \cdot r$ we determine the error correction for the next learning cycle. As the algorithm is unsupervised, the weights are adjusted depending on the values of the previous iteration. For the first iteration Δw_{ij} = 7.453 · .05 resulting in a weights adjustment of 0.373. Consequently w1 = .174 + .373 = .547 and the w2 = .092 + .373 = 0.465 for the next iteration. This weights adjustment process is repeated every iteration.

$$\Delta w_{ij} = c \cdot r \tag{5}$$

Orientation	**W1**	**W2**	**Threshold**	**Iteration**	**Throughput**	**Slope**	**Error Correction**	**Learning Rate**	**Suggested orientation**
				0	0				
40	0.17392098	0.092288451		1	7.45344	7.45344	0.372672	0.05	-70
-70	0.54659298	0.464960451	1	2	2.89576	-4.55768	-0.227884	0.05	-40
-40	0.31870898	0.237076451	1	3	7.45344	4.55768	0.227884	0.05	-50
-50	0.54659298	0.464960451	1	4	5.99287	-1.46057	-0.0730285	0.05	-20
-20	0.47356448	0.391931951	1	5	9.05292	3.06005	0.1530025	0.05	-10
-10	0.62656698	0.544934451	1	6	9.56064	0.50772	0.025386	0.05	10
10	0.65195298	0.570320451	1	7	9.56064	0	0	0.05	20
20	0.65195298	0.570320451	1	8	9.05292	-0.50772	-0.025386	0.05	20
20	0.62656698	0.544934451	1	9	9.05292	2.51E-15	1.26E-16	0.05	10
10	0.62656698	0.544934451	1	10	9.56064	0.50772	0.025386	0.05	10
10	0.65195298	0.570320451	1	11	9.56064	0	0	0.05	20

Table 3: Learning process with 0.05 learning constant

Table 3 illustrates the learning cycle. It shows that when the throughput is increasing also the weights for the next iteration are increasing. On the other hand, when the throughput is decreasing the weights also. When the error correction is 0, it means that is near of stabilize the system as the iteration 7 shows. Beside, the maximum scale of error correction is bounded by the maximum achievable 9.561 · .05 = .478. There is no supervision in this algorithm, neither random changes in any iteration (e.g. there is no weight adjustment with random values). At the end, the optimal weight and orientation are

shown. Table 1 illustrates that the best throughput is 9.561 and for that, the system can choose from -10º to 10º in the antenna orientation. Furthermore, -20º and 20º are also acceptable with a throughput of 9.053. As a result of this, the algorithm was in the optimal values since iteration 5 and in the iteration 6 it arrives until the best one. We show until the iteration 11 to demonstrate that the system was stabilizing and converging in optimal values.

Finally, the optimal values are w1 = .652 , w2 = .570 and throughput = 9.561.

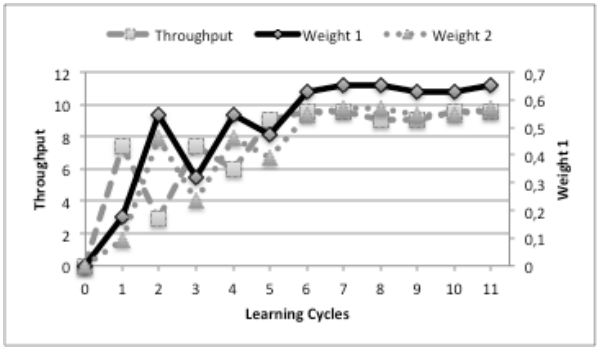


Figure 2: Learning cycles of throughput, weight 1 and weight 2

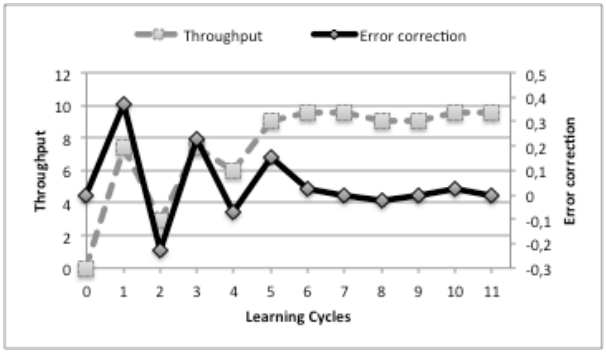


Figure 3: Learning cycles of throughput versus error correction

Figure 2 illustrates the progression of the learning cycles and the relationship between synaptic weights and throughput. Figure 3 illustrates the relationship between throughput and error correction. Figures 2 and 3 illustrate an undershoot/overshot period in which throughput varies greatly (cycles 1-5). As the cycles progress stabilisation of learning and optimisation of collective throughput is achieved (cycles 6-11).

Figure 4 illustrates the relationship between linear regression of previous throughput (slope) and throughput. Figure 5 illustrates the relationship between antenna orientation and throughput.

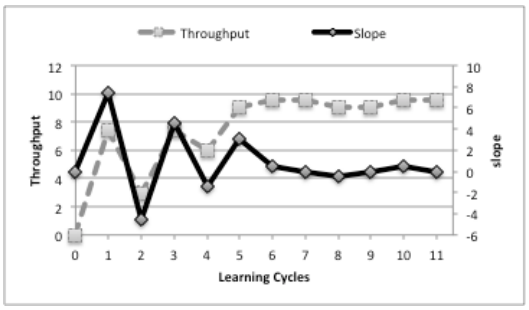


Figure 4: Learning cycles of throughput versus slope

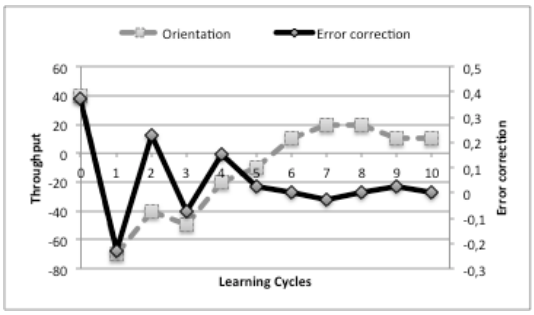


Figure 5: Learning cycles of orientation versus error correction

Figures 2-5 illustrate that initially the system is learning and trying different possible values. After iteration 5 the PLAAO algorithm enters an optimisation phase in which collective throughput improves gradually from cycle to cycle.

5 Conclusions and Future Work

In this work we proposed PLAAO, a feed forward neural network which utilises directed learning. PLAAO is a SON approach, which uses historic antenna orientation to maximize throughput for all MNs in the coverage of a base station. Our algorithm is implemented as an extensible approach in the

R language. The approach considers a heterogeneous network scenario with multiple cell propagation from a radio base station. Using system simulations (NS-3) of a sample LTE, we verify that the system is learning. Results illustrate improved collective throughput utilising the PLAAO algorithm. Future work will focus on improving the learning efficiency of the PLAAO algorithm to achieve optimal collective throughput regardless of the initial random weight selection.

References

[1] http://www.ericsson.com/news/1561267

[2] Schmelz, L.C.; Amirijoo, M.; Eisenblaetter, A.; Litjens, R.; Neuland, M.; Turk, J., "A coordination framework for self-organisation in LTE networks," *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on* , vol., no., pp.193,200, 23-27 May 2011

[3] Bandh, T.; Romeikat, R.; Sanneck, H.; Haitao Tang, "Policy-based coordination and management of SON functions," *Integrated Network Management (IM), 2011 IFIP/IEEE International Symposium on* , vol., no., pp.827,840, 23-27 May 2011

[4] Sinclair, N.; Harle, D.; Glover, I.A.; Irvine, J.; Atkinson, R.C., "An Advanced SOM Algorithm Applied to Handover Management Within LTE," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1883,1894, Jun 2013

[5] Munoz, P.; Barco, R.; de la Bandera, I., "On the Potential of Handover Parameter Optimization for Self-Organizing Networks," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1895,1905, Jun 2013

[6] Awada, A.; Wegmann, B.; Viering, I.; Klein, A., "A SON-Based Algorithm for the Optimization of Inter-RAT Handover Parameters," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1906,1923, Jun 2013

[7] Hongcheng Zhuang; Shmelkin, D.; Zezhou Luo; Pikhletsky, M.; Khafizov, F., "Dynamic Spectrum Management for Intercell Interference Coordination in LTE Networks Based on Traffic Patterns," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1924,1934, Jun 2013

[8] Maso, M.; Debbah, M.; Vangelista, L., "A Distributed Approach to Interference Alignment in OFDM-Based Two-Tiered Networks," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1935,1949, Jun 2013

[9] de Lima, C.H.M.; Bennis, M.; Latva-aho, M., "Statistical Analysis of Self-Organizing Networks With Biased Cell Association and Interference Avoidance," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1950,1961, Jun 2013

[10] Munoz, P.; Barco, R.; Ruiz-Aviles, J.M.; de la Bandera, I.; Aguilar, A., "Fuzzy Rule-Based Reinforcement Learning for Load Balancing Techniques in Enterprise LTE Femtocells," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1962,1973, Jun 2013

[11] Fehske, A.J.; Klessig, H.; Voigt, J.; Fettweis, G.P., "Concurrent Load-Aware Adjustment of User Association and Antenna Tilts in Self-Organizing Radio Networks," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1974,1988, Jun 2013

[12] Engels, A.; Reyer, M.; Xiang Xu; Mathar, R.; Jietao Zhang; Hongcheng Zhuang, "Autonomous Self-Optimization of Coverage and Capacity in LTE Cellular Networks," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.1989,2004, Jun 2013

[13] Toril, M.; Luna-Ramirez, S.; Wille, V., "Automatic Replanning of Tracking Areas in Cellular Networks," *Vehicular Technology, IEEE Transactions on* , vol.62, no.5, pp.2005,2013, Jun 2013

[14] D. Paetridge and K.M. Hussain. “Knowledge Based Information Systems”. McGraw Hill, 1994

[15] S. Sun, W. Wang, M. Wei and Z. Zhou, “Load Balancing Based on Subscriber Characteristic for MSC POOL System,” Computational and Information Sciences (ICCIS), 2012 Fourth International Conference on , vol., no., pp.1111,1115, 17-19 Aug. 2012.

[16] N. Sinclair, D. Harle, I. Glover, J. Irvine and R. Atkinson, “An Advanced SOM Algorithm Applied to Handover Management Within LTE,” Vehicular Technology, IEEE Transactions on , vol.62, no.5, pp.1883,1894, Jun 2013

[17] W. Culloch, W Pitts, “A logical calculus of the ideas immanent in nervous activity.

[18] D. Hebb, “The Organization of Behaviour”, published by Wiley.

[19] Weisi Guo; Siyi Wang; Yue Wu; Rigelsford, J.; Xiaoli Chu; O'Farrell, T, “Spectral- and energy-efficient antenna tilting in a HetNet using reinforcement learning,” Wireless Communications and Networking Conference (WCNC), 2013 IEEE , vol., no., pp.767,772, 7-10 April 2013.

Microstrip Line Fed Patch Antenna suitable for WBAN Applications.

Senan Morris[1], Nick Timmons[1], *Member IEEE* and Jim Morrison[1], *Member IEEE*

[1]WiSAR Lab, LYIT, Letterkenny, Co. Donegal, Ireland

senanmorris@hotmail.com

Nick.Timmons@lyit.ie

Jim.Morrison@lyit.ie

Abstract

Over the past number of years there has been a huge increase in the study of wireless body area networks (WBAN). Many proposed applications of BAN technology are in the medical field including body worn wireless sensors used to monitor an individual's health. If these devices are to be accepted by the majority of consumers, the hardware including the radio transceiver and especially the antenna need to be low profile, unobtrusive, and comfortable. A micro-strip feed line patch antenna design is presented which offers an ideal solution due to its low profile and off-body radiation characteristics. The prototype is fabricated on an FR4 substrate with a dielectric constant of 4.4 and thickness of 1.6mm. The antenna, operating at the IEEE 802.15.6 BAN standard frequency of 2.45GHz, is highly suitable for BAN applications, as the device can be can be worn flat on the surface of the body, or in clothing. Also the ground plane shields the antenna from the desensitising effects of the human body which is an essential characteristic for a body worn antenna.

1 Introduction

Wireless body area network (WBAN) is the term used to describe a network of devices connected wirelessly for communication on, in or near the human body. WBAN applications appear in many applications such as gaming and sports, but the main application of the WBAN is in health care. As the average life expectancy and the earth's population are always increasing, this has caused a significant rise in healthcare costs in many countries and is the primary motivating factor for innovation in health care. These health monitoring devices may include devices that are able to tell whether someone is about to have a heart attack by monitoring their vital signs and alerting the patient that they need to receive medical attention urgently. Others may include a device that is able to monitor a diabetics insulin levels and if their levels drop, a pump will automatically inject them with insulin. There are a lot of problems that arise when designing these devices, and the main problem concerned with consumer acceptance of the WBAN is size and visibility of the device. For these devices to be accepted by the majority of consumers, the radio system components, including the antenna need to be somehow hidden, compact and low weight.

Microstrip antennas can be traced back to the early 1950's but didn't gain any considerable attention until the 1970's. Nowadays the microstrip antenna is used in many applications such as GPS systems and mobile satellite communications [1] - [4]. Patch antennas have also recently begun to show promising results for body worn communications [5]–[6]. The reason that the microstrip antenna is desirable for these applications is because of its low cost, light weight, planar structure and also its ease of fabrication. The microstrip antenna can also be easily integrated with microwave integrated circuitry and is capable of dual and triple frequency operations. But the patch antenna also has its disadvantages, such as narrow bandwidth, low efficiency and gain and also has a low power handling capacity [8].

Nowadays extensive research and development of the microstrip has been aimed at exploiting their advantages.

In paper [7], a microstrip rectangular patch antenna for wearable applications operating from 1710-1785MHz and 1805MHz is presented. A patch antenna was chosen for the design because of their low profile and also the fact that the ground plane of the patch antenna effectively shields the antenna from the influence of the human body and the user from the negative effects of the electromagnetic field produced by the antenna. The copper radiating elements of the antenna are fabricated onto a substrate with a dielectric constant (ε_r) of 2.6 and a height of 1.524mm. The antenna is fed by means of a microstrip transmission line and includes a ¼ wave transformer in order to match the transmission line impedance to the antenna impedance. Two opposite corners of the patch were truncated in order to improve S11 parameters.

The simulated reflection co-efficient was found to be in good agreement with that of the measured reflection co-efficient. Both were found to be resonating at the intended frequency of 1.79GHz. The reflection co-efficient was then measured in close proximity to the human body. The antenna was placed on the uniform of a human volunteer. There was little or no change to the reflection co-efficient when the antenna was placed on the human body when compared to the free space measurement. The radiation patterns were measured in an anechoic chamber. It was firstly measured in free space and then measured on the uniform of a human volunteer. Once again there was little or no change to the free space pattern and the pattern when measured on the human body. This led to the conclusion, that the ground plane had indeed shielded the antenna from the effects of the human body. This also led to the main motivation for the design used in this paper.

The next section of this work will discuss the equations used in the mathematical design of the antenna. Section 3 will then discuss the experiments and measurements conducted on the antenna. The fourth section will illustrate and discuss the results obtained from the experiments discussed in section 3. And finally section 5 will contain the conclusion from this work and any proposed future work.

2 Antenna Design

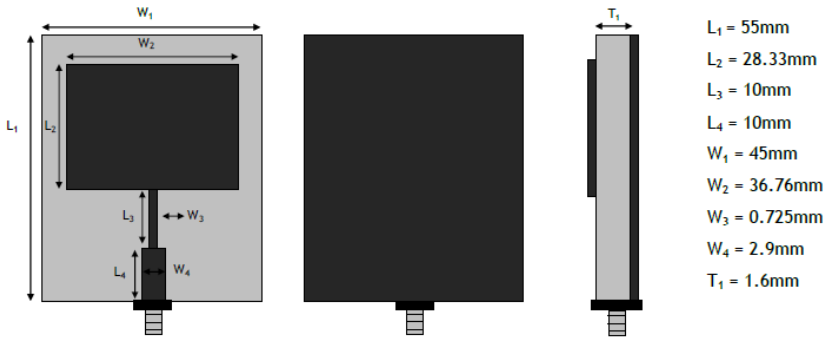


Fig.1 Patch Antenna Geometry

Fig.1 illustrates the geometry of the antenna. The copper radiating element and ground plane of the antenna are fabricated onto a FR4 substrate (Dielectric constant $\varepsilon_r = 4.4$, thickness = 1.6mm). The design of a rectangular patch antenna depends mainly on three parameters: the dielectric constant of the substrate, the substrate thickness and the intended operating frequency. All equations used in the mathematical design of the antenna are listed in Fig.2. Equations (1) – (5), taken from [8], are the equations that were used in the mathematical design of the copper radiating patch element of the antenna. The first calculation that needs to be made is the width of the copper patch (W_2 from Fig.1), which can be calculated from equation (1). Equations (2-4) are steps used to calculate the actual length

of the copper radiating patch. So the next calculation that needed to be made in order to calculate the actual length is the effective dielectric constant (2). This calculation needs to be made because the electric field lines move through the air before entering the dielectric substrate [9]. This value is a slightly lower value than the dielectric constant of the substrate. The third calculation that needs to be made is the effective length of the patch (3). After calculating the effective length of the patch, the delta length then needs to be calculated from equation (4). The actual length can then be calculated from equation (5) which was the equation used to find L_2 in Fig.1. Equations (6, 7) are used to calculate the minimum requirements for the length and width of the ground plane. These equations relate to L_1 and L_2, respectively, from Fig.1. These calculation (6, 7) had to be significantly adjusted in order to feed the antenna with a microstrip feed line.

As the antenna is fed by means of a microstrip feed line, this meant that it must match the impedance of the transmission line. In a coaxial cable, which is the most commonly used in antenna applications and also used in this paper, the transmission line impedance is standardized to 50Ω. It can also be standardized to 75Ω but that impedance is mainly used for satellite communications. Equation (8) was used to calculate the width of the microstrip feed line (W_4 in Fig.1). As Z_0 = 50Ω and h=1.6mm, the equation was then transposed in order to solve for *w*. Finally a quarter wavelength transformer was used to match the transmission line impedance (50Ω) to the high input impedance (243Ω) of the patch antenna. If the input impedance does not match the transmission line impedance, this would result in a mismatch causing a lot of the power to be reflected back to the source and potentially damaging it. Using a quarter wavelength transformer does not require the use of any extra matching circuitry which is another benefit of this design.

Width:

$$w = \frac{C}{2.fc\sqrt{\frac{(Er+1)}{2}}} \quad (1)$$

Where:
c = Speed of light ($3x10^8$)
fc = Operating Frequency (2.45GHz)
Er = Dielectric Constant of Substrate (4.4)

Effective Dielectric Constant:

$$E_{reff} = \frac{E_r+1}{2} + \frac{E_r-1}{2}[1 + 12\frac{h}{w}]^{-\frac{1}{2}} \quad (2)$$

Where:
h = height/thickness of dielectric Substrate
w = width of the radiating patch

Effective Length:

$$L_{eff} = \frac{c}{2.f_c.\sqrt{E_{reff}}} \quad (3)$$

Delta Length:

$$\Delta L = 0.412.h.\frac{(E_{reff}+0.3)\left(\frac{w}{h}+0.264\right)}{(E_{reff}-0.258)\left(\frac{w}{h}+0.8\right)} \quad (4)$$

Actual Length:

$$L = L_{eff} - 2.\Delta L \quad (5)$$

Ground Plane Minimum:

$$Lg(length) = 6.h + L \quad (6)$$

$$Wg(width) = 6.h + w \quad (7)$$

Where:
h = height/thickness of substrate
w = width of copper patch
L = actual length of copper patch

Characteristic Impedance of a Microstrip Feed Line:

$$Z_0 = \frac{60}{\sqrt{E_{reff}}}.\ln(\frac{8h}{w} + 0.25\frac{w}{h}) \quad (8)$$

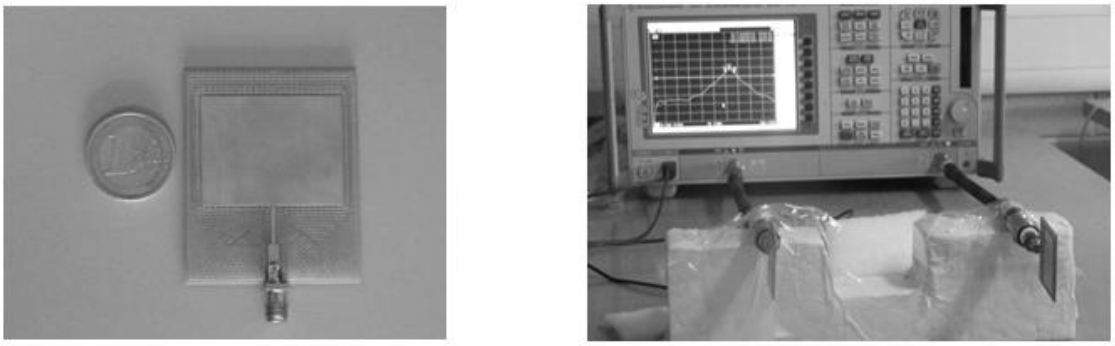
Where:
Z_0 = Characteristic Impedance
w = width of microstrip feed line
h = height /thickness of substrate (1.6mm)

Fig.2 Design Equations for Microstrip Line Fed Patch Antenna

3 Measurements

To test the performance of the antenna there are various test measurements that may be carried out on the device. Two standard and common antenna test measurements are the $\boldsymbol{S_{11}}$ and $\boldsymbol{S_{21}}$ measurements. The $\boldsymbol{S_{11}}$ or reflection co-efficient test measurement is used to see how well the antenna impedance is matched to the transmission line impedance whereas the $\boldsymbol{S_{21}}$ is a measurement of the antenna's forward voltage gain. Both measurements can be represented graphically and are measured in dB. An $\boldsymbol{S_{11}}$ or reflection co-efficient graph is a graphical representation of the power reflected back from the antenna in the transmission line. As previously discussed a poorly matched antenna or antenna with a poor return loss can result in poor antenna performance. In a worst case scenario a complete mismatch could result in damaging the power source. The return loss of an antenna is normally measured at the centre operating frequency of the antenna. -18dB is an industry standard for antenna return loss measurements. The $\boldsymbol{S_{21}}$ measurement is used to measure the peak gain or power radiated from the antenna. Poor gain can result in poor performance and end up with the antenna having a poor range.

Simulation measurements were carried out first via the use of Ansoft's HFSS (High Frequency Structure Simulator) software [10]. The mathematically calculated dimensions of the antenna from section 2 were inputted into the software. Some fine tuning on the antenna's dimensions had to be completed in order to obtain satisfactory simulation results. Once they had been achieved, the antenna was then manufactured. Real life test measurements were then performed on the antenna via the use of a Rode & Shwarz ZVB 8 Vector Network Analyser (VNA). It should be noted that the only recorded simulation measurement was the free space $\boldsymbol{S_{11}}$ measurement. All other test results were obtained from real life test measurements on the antenna via the VNA.



(a) (b)

Fig.3 (a) Manufactured Patch Antenna (b) Test measurement set-up

4 Results & Discussion

4.1 Return Loss Measurements $(\boldsymbol{S_{11}})$

Shown in Fig.4 are the measured and simulated reflection co-efficient, which were found to be in good agreement. The test measurement found the antenna to be resonating from 2.41 – 2.48GHz with a centre

frequency of 2.45GHz. At the centre frequency the return loss was -22.81dB. As the antenna was designed for BAN applications the same measurements were also recorded on body to see how it would affect the antenna performance. Fig.5(a) and 5(b) show a comparison of each measurement that was taken when the antenna was positioned on the upper arm and the shoulder respectively. In each of these positions, the antenna was firstly placed (ground side) directly onto the skin of the body, then 5mm, 10mm and 15mm respectively, away from the body. From Fig.5(a) and 5(b) it is evident that when the antenna came into direct contact with the human body, the Return Loss measurement decreased when compared to the free space measurement. The decrease in the Return Loss measurement is caused by the human body absorbing some of the energy transmitted to the antenna from the source. Even with the antenna placed at distances of 5mm and 10mm away from the body, it was still absorbing some of the power and having a minor effect on the Return Loss measurement. When the antenna was placed at a distance of 15mm away from the body, it had little or no effect on the Return Loss measurement. It should be noted that more energy was absorbed by "fattier" parts of the body.

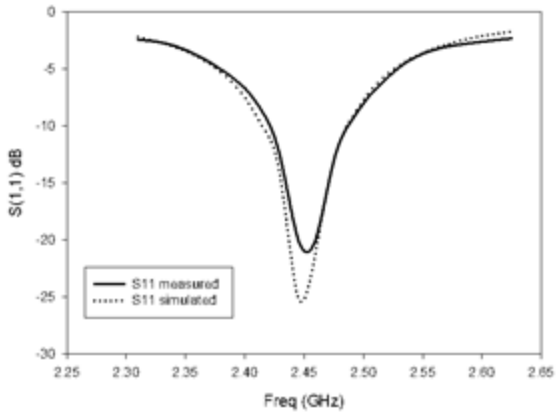
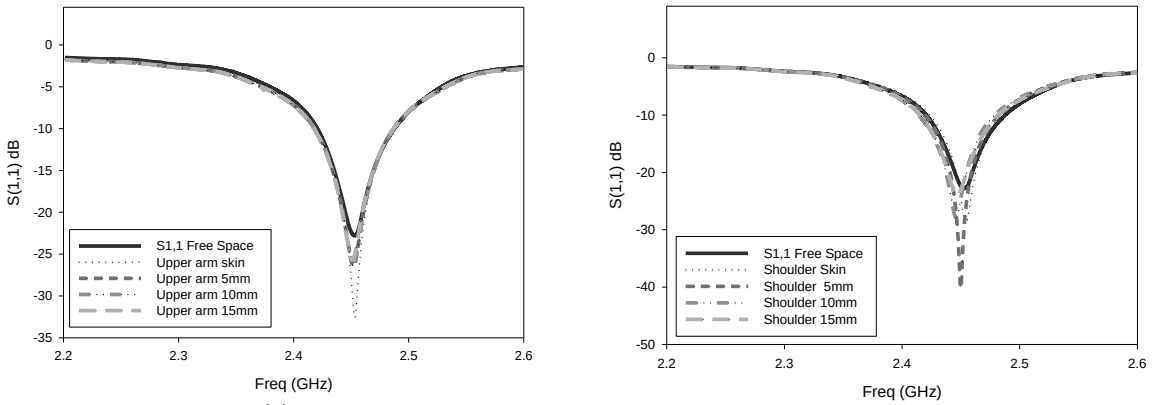
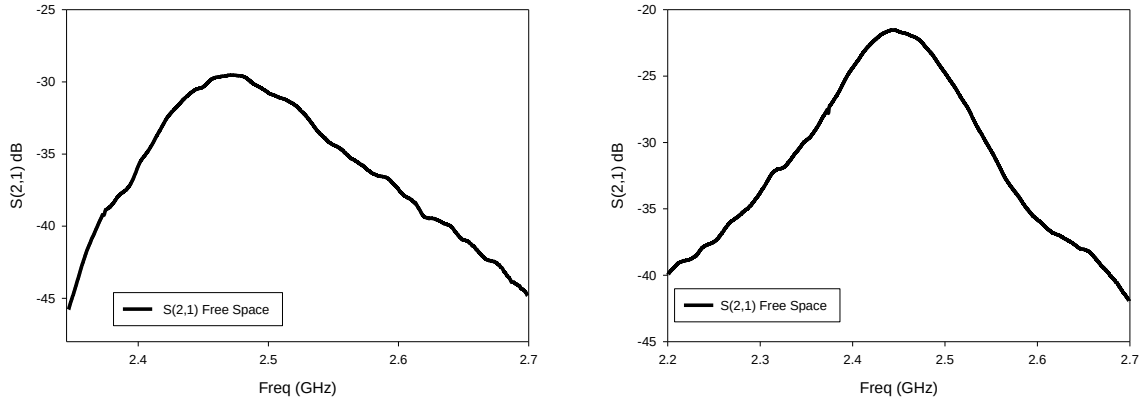


Fig.4 Simulated and Measured reflection co-efficient



(a) (b)

Fig.5 (a) Measured reflection co-efficient with antenna positioned on the upper arm (b) Measured reflection co-efficient with antenna positioned on the shoulder

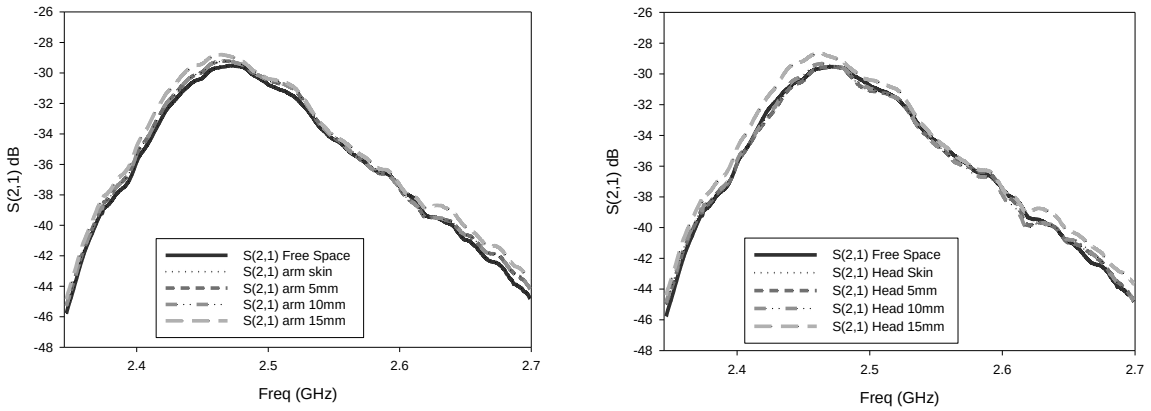


(a) (b)

Fig.6 (a) Measured S_{21} in Free Space (b) "Two Antenna Method" S_{21} measurement before factoring out free space loss

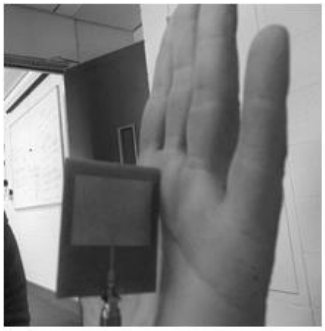
3.2 Gain Measurements, (S_{21})

In addition to the Return Loss measurements, the S_{21} gain measurements were also recorded. The first measurements were recorded with the use of a second non-identical antenna which was resonating at the same frequency as the AUT (antenna under test), 2.4 - 2.48GHz. Shown in Fig.6 (a) is the S_{21} measurement that was recorded in free space. The antenna was found to have peak radiation from 2.42-2.53GHz, which shows that proposed antenna operates within the 802.15.6 2.45GHz frequency band. As with the S_{11} measurements, the S_{21} measurements were also recorded in various positions around the body in order to see how the body would affect the antenna's performance. S_{21} measurements were also recorded with the antenna placed directly onto the skin of the body and then 5mm, 10mm and 15mm away from the body. Fig.7(a) and 7(b) show a comparison of each measurement taken when the antenna was positioned on the arm and on the head. It is evident from these measurements that when the antenna was placed flat on the body and placed short distances away from the body, that the human body had little or no effect on the antenna's performance. The S_{21} measurement was then recorded with a hand placed vertically against the side of the antenna, which is illustrated in Fig 8(a).

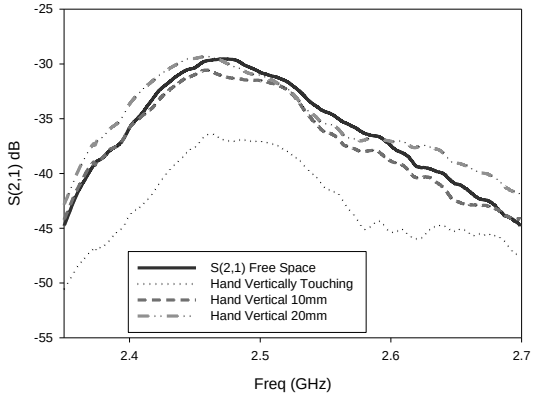


(a) (b)

Fig.7 (a) Measured S_{21} with antenna positioned on arm (b) Measured S_{21} with antenna positioned on head



(a)



(b)

Fig.8 (a) Hand placed vertically against the side of antenna (b) Measured $\boldsymbol{S_{21}}$ with hand placed vertically against the side of the antenna

Fig.8(b) shows a comparison of the measurements taken with a hand placed vertically against the side of the antenna. The measurement shows that when the hand came into contact with the antenna, it dramatically affected the gain of the antenna. This was due to the fact the hand was absorbing some of the radiated power from the antenna. The measurements were also recorded with the hand positioned 10mm and 20mm away from the antenna which appeared to have little or no effect on the antenna gain.

The antenna was found to have an actual gain of 2.52dB. Fig.6(b) shows the measurement before factoring out free space loss. The gain was measured by using the "Two Antenna Method" [9], which requires the use of a second identical antenna separated by a distance, which must satisfy the far-field criterion of the antenna. When power is radiated from an antenna, the radiated waves cross the near field and field regions. The far field ranges from 2λ to infinity.

3.3 Discussion

The patch antenna displays reasonable performance when compared to the performance of a ¼ wavelength monopole antenna. The ¼ wavelength monopole is a standard and common antenna design. The design consists of a straight rod like conductor, in most cases a piece of wire, mounted perpendicularly over a ground plane. The antenna has a standard and well known gain of 1-2dB which is lower than the gain of the patch antenna presented in this paper. It also has an omnidirectional radiation pattern which differs to the radiation pattern of the patch antenna. Although not measured in this paper, the patch antenna has a directional radiation pattern which makes the antenna ideal for off-body communication. The patch antenna can be placed flat on the surface or skin of the human body and radiates away from the body. The monopole cannot be placed flat on the surface of the human body and would be quite obtrusive. It does not radiate away from the body which allows the body to absorb the power being radiated from the monopole.

4 Conclusion & Future Work

A microstrip line fed patch antenna with quarter wavelength-transformer is presented. The patch antenna was found to be resonating from 2.4-2.48GHz with a centre frequency of 2.454GHz. At the centre frequency the return loss was –22.8dB. When mounted flat onto the skin of the human body the body absorbed power causing less power to be reflected back to the source. The antenna was found to be radiating at the intended frequency band of 2.4-2.48GHz with a measured gain of 2.52dB. Moreover the presence of the human body had little effect on the antenna's performance, due to the shielding effect from the ground plane. This together with its low profile, the fact it can be placed directly onto the surface of the human body and that it radiates away from the body makes the antenna highly suitable for off- body BAN applications.

An extension of this work would be to design a flexible antenna. As the antenna presented in this paper is made from a rigid FR4 substrate it could possibly affect the wearer's comfort. The aim would be design to design a truly flexible antenna (spring back to its original shape) that can still maintain a reasonable performance when being flexed or bent. There are many design considerations to be taken into account when designing a flexible antenna such as suitable substrate choice and conductive material. Further research is currently being conducted on flexible antennas.

5 References

[1] Hua-Ming Chen, Yang-Kai Wang, Yi-Fang Lin, Che-Yen Lin, Shan-Cheng Pan, *Microstrip-Fed Circularly Polarized Square-Ring Patch Antenna for GPS Applications*. IEEE Transactions on Antennas and Propagation, 2009

[2] George B. Abdelsayed, *Triple-band circularly polarized slotted patch antenna for GPS and UMTS systems*. German University in Cairo (GUC) Cairo, Egypt, 2010.

[3] Radha Telikepalli, *Design of a Wide Band Microstrip For use in a Phased Array Antenna for Mobile Satellite Communications.* Ottawa, 1995

[4] Josaphat Tetuko Sri Sumantyo, Koichi Ito, and Masaharu Takahashi, *Dual-Band Circularly Polarized Equilateral Triangular-Patch Array Antenna for Mobile Satellite Communications*. IEEE Transactions on Antennas and Propagation, 2005

[5] Gareth A. Conway and William G. Scanlon, *Antennas for Over-Body-Surface Communication at 2.45 GHz*. IEEE Transactions on Antennas and Propagation, 2009

[6] Gareth A. Conway, William G. Scanlon, and D. Linton, *Low Profile Microstrip Patch Antenna for Over-Body Surface Communication at 2.45GHz*. Technology Conference, VTC2007-Spring.

[7] Mariusz Wozniak, Adrian Durka, Mariek Bugaj, Rafal Przesmycki, Leszek Nowosielski and Marian Dunk, *Designing and optimization of a Microstrip Rectangular Patch Antenna to work on the Human Body. ", Mikon 2012, 19th international conference on Microwaves, Radar and Wireless Communication, May 21-23, Poland*

[8] Z.I. Dafalla, W.T.Y. Kuan, A.M. Abdel Raman and S.C. Shudakar, *Design of a Rectangular Microstrip Patch Antenna at 1GHz*. RF and Microwave conference, 2004

[9] C.A Balainis, *Antenna Theory, Analysis and Design*. 3rd Edition, Wiley, 2005

[10] Ansoft High Frequency Structure Simulator v12

ISSN 1649-1246