IT&T

The IT&T
12th International Conference on Information Technology and Telecommunication
2013

Small Devices, Big Data, Real Challenges

Instítiúid Teicneolaíochta
Bhaile Átha Luain
Athlone Institute
of Technology

9 - 10 May 2013

Editors: Markus Hofmann & Enda Fallon

ISSN 1649-1246

IT&T 2013 General Chair’s Letter

As the Chair of the 12th Information Technology and Telecommunications Conference (IT&T 2013), I have great pleasure in welcoming you to this year’s conference which will be hosted by the School of Engineering and Software Research Institute (SRI) at Athlone Institute of Technology on the 9th and 10th of May 2013.

This year’s conference theme is “Small Devices, Big Data Real Challenges”. 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. The effective creation, transmission and analysis of such data provides real challenges for the research and development community.

This year’s conference includes 16 research papers and 11 posters from research from Irish Institutes of technology and Universities. The papers and posters will be presented six sessions. The Key-note speech will be delivered by Matt Hamilton from Ericsson. Matt works in the area of telecommunications network management and will provide an overview of Ericsson’s research and development focus in this space.

Many people have helped with the preparation of this year’s conference. First and foremost I would like to thank Dr Markus Hofmann of IT Blanchardstown. As General IT&T Committee Chair Markus has provided invaluable support to this year’s conference. Thanks are also due to Dr Nick Timmons of Letterkenny IT for his considerable and on-going support of the IT&T conference. Special thanks to the technical programme committee for their timely and considered feedback on submitted material. Thanks also to Institutes of Technology Ireland (IOTI) and Ericsson for their generous sponsorship of this year’s event. Within AIT, special thanks to the staff of the School of Engineering (Dr Austin Hanley), Software Research Institute (Dr Brian Lee, Anthony Cunningham and Dr Yuansong Qiao), Office of Research (Paul Killeen, Lorna Walsh and Stephanie Lennon) for helping with all the local arrangements and the conference registration.

A thank you also to the session chairs for helping run the conference programme.

I herewith welcome you to Athlone and to the IT&T 2013 conference.

Dr Enda Fallon
Local Chair, IT&T 2013
Software Research Institute
Athlone Institute of Technology
Athlone, Ireland

Technical Programme Committee Chair’s Letter

Dear Colleagues,

As Technical Programme Chair, I would like to welcome you to the Twelfth Information Technology and Telecommunications Conference (IT&T 2013) hosted by the Athlone Institute of Technology, 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 Twelfth IT&T is “Small Devices, Big Data, Real Challenges”. There is little doubt the devices that are now used are smaller and smarter than 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 Big Data. 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 I express our sincere gratitude to all of them for their help in the reviewing process. The outcome of the review process produced sixteen papers that were accepted and these will be presented during the five technical sessions spanning the two days of the conference. This year’s conference will also display a number of posters.

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

Prof. John Murphy, University College Dublin, Ireland

IT&T 2013 Chairs and Committees

Local Chair

Enda Fallon, Athlone Institute of Technology

IT&T General Chair

Markus Hofmann, Institute of Technology Blanchardstown

Technical Programme Committee Chair

John Murphy, University College Dublin

Patronage & Sponsor Chair

Nick Timmons, Letterkenny Institute of Technology

Submissions Chair

Brian Lee, Athlone Institute of Technology

Publicity Chair

Anthony Cunningham, Athlone Institute of Technology

Organising Committee

Brian Nolan, Institute of Technology Blanchardstown
Enda Fallon, Athlone Institute of Technology
Gabriel-Miro Muntean, Dublin City University
John Murphy, University College Dublin
Markus Hofmann, Institute of Technology Blanchardstown
Nick Timmons, Letterkenny Institute of Technology
Paul Doyle, Dublin Institute of Technology
Phelim Murnion, Galway-Mayo Institute of Technology
Brett Becker, College of Computer Training
Arnold Hensman, Institute of Technology Blanchardstown

Technical Programme Committee

Enda Fallon, AIT
Austin Hanley, AIT
Brian Lee, AIT
Anthony Cunningham, AIT
Yuansong Qiao, AIT
Shane Banks, AIT
Sheila Fallon, AIT
Paul Archbold, AIT
Ronan Flynn, AIT
Dirk Pesch, CIT
Nick Timmons, LYIT
Declan O'Sullivan, TCD
Ken Brown, UCC
Tom Pfeifer, W-ITC
Dave Denieffe, ITC
Markus Hofmann, ITB
Matt Smith, ITB
Larry McNutt, ITB
Brian Nolan, ITB
Gabriel-Miro Muntean, DCU
John Murphy, UCD
Antonio Ruzzelli, UCD
David Tracey, Dell
Martin Klepal, CIT
John Murphy, UCD
Jim Morrison, LYIT
Cristina Muntean, NCI
Cormac J. Sreenan, UCC
Sean McGrath, UL
Martin McGinnity, UU
Mark Riordan, IADT
Kieran Delaney, CIT
Paul Walsh, CIT
Jim Clarke, TSSG
Liam Kilmartin, NUIG
Ian Pitt, UCC
Pat Coman, ITT
Mark Davis, DIT
Dave Lewis, TCD
Nigel Whyte, ITC
Marcus Keane, NUIG
Gregory Provan, UCC
Paul Doyle, DIT
Phelim Murnion, GMIT
Gabriel Hogan, Ericsson
Arnold Hensman, ITB
Sidath Handurukande, Ericsson

Table of Contents

Session 1: Networking I

Session 2: Information Analysis and Management

Session 3: Text and Data Mining

Session 4: Web and Cloud Technologies

Session 5: Networking II

Short Papers - Poster Session:

ITT13 Author Index

Session 1

Networking I

Device-Oriented Energy-Aware Utility-Based Priority Scheduler for Video Streaming over LTE System

Longhao Zou, Ramona Trestian, Gabriel-Miro Muntean
Performance Engineering Laboratory, Dublin City University, Ireland
longhao.zou3@mail.dcu.ie, ramona@eeng.dcu.ie, munteang@eeng.dcu.ie

Abstract

Nowadays people tend to spend most of their time in front of a screen, and expect to be able to connect to the Internet anytime and anywhere and from any type of mobile device. Therefore, fast surfing speed on Internet, high resolution display screen, advanced multi-core processor and lasting battery support are becoming the significant standards in the nowadays mobile devices. In this context the network operators must be able to differentiate between their multiscreen offerings in order to ensure uninterrupted, continuous, and smooth video streaming with minimal delay, jitter, and packet loss. This paper proposes a novel Device-Oriented Energy-Aware Utility-based Priority scheduling (DE-UPS) algorithm which makes use of device differentiation in order to ensure seamless multimedia services over LTE networks. The priority decision is based on the device classification, energy consumption of the mobile device and the multimedia stream tolerance to packet loss ratio.

Keywords: Long-term Evolution, Scheduling Algorithm, Utility Functions, Energy Consumption, Quality of Service

1 Introduction

The increasing demand for massive data network consumption, such as music streaming, video streaming, social networking, live gaming, navigation and cloud sync, with the limitation of Quality of Service (QoS) requirements, puts pressure on the next generation mobile networks. The Long Term Evolution (LTE), an evolution of the GSM/UMTS standards which aims to design the all-IP network architecture highly improves the spectrum efficiency and significantly reduces the transfer latency. However, the main challenge that the mobile network operators are facing is the ability to differentiate between the multiscreen offerings in order to provide seamless multimedia experience with minimal delay, jitter, and packet loss, to their customers.

Consequently, in this paper we propose a novel priority-based scheduling technique for multimedia streaming over LTE networks. The proposed scheduler takes into account the QoS constraints of the multimedia application, and efficiently utilizes the information about the device display resolution and energy consumption in order to prioritize the resource allocation and ensure the best multimedia experience to the mobile users.

2 Related Works

A Maximum Sum Rate (MSR) scheduling scheme without transmission power adaption was presented in [1], well suited to the limited dynamic power range downlink scenario. However, in the case of unfair sharing of the radio resources and having latency requirements, such scheduling methods are unsuitable. Delay aware downlink scheduling schemes in OFDMA systems are proposed in [2] [3]. These schemes select the highest priority to the user based on channel conditions and the amount of queuing delay for real-time or non-real-time services. Moreover, a q-learning based scheduling scheme proposed in [4] enables fair provision of different throughput in terms of the different classes of users.

Another challenge is the multimedia service delivery with QoS provisioning over the wireless environment where connections are prone to interference, high data loss rates, and/or disconnections. In this context there has been extensive academic research related to adaptation techniques for video streaming especially over wireless networks. Various solutions have been proposed in the literature [5]-[8] that address this problem of streaming video over the Internet while maintaining high end-user perceived quality levels and make efficient use of the wireless network resources.

However, most of the previous works do not consider the attributes of the devices used at the end-user side. The resolution of device display tends to be higher and higher and the limitation of lifetime of battery restricts long-term working of mobile devices. Therefore, this paper proposes a utility-based priority scheduler based on utilities related to resolution of device display, device energy consumption and estimated QoS requirements of the transmitted video stream.

2 Device-Oriented Energy-Aware Utility-Based Scheduling Scheme

2.1 Resource Allocation Strategy in LTE

In the downlink transmission, an efficient time-frequency modulation technology is exploited, namely Orthogonal Frequency-Division Multiple Access (OFDMA). The unit of OFDMA named Resource Block (RB) contains 12 consecutive subcarriers of 180 kHz bandwidth in the frequency domain, and in the time domain it accounts 0.5 millisecond time slot [9]. Two consecutive RBs referred to as Physical Resource Block (PRB) are assigned to a user for a Transmission Time Interval (1 millisecond). A brief illustration of resource allocation is shown in Figure 1. Considering a number N of UEs competing for resources, by using a scheduler function, each UE will get allocated PRBs on the physical channel in the time-frequency domain based on some specified conditions, such as channel states, QoS requirements or fairness conditions.

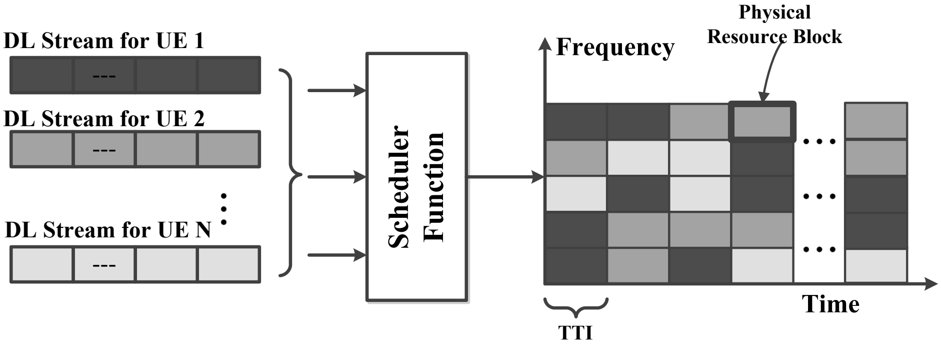


Figure 1. A brief Description of Resource Allocation

2.2 Scheduling Scheme Utility Function

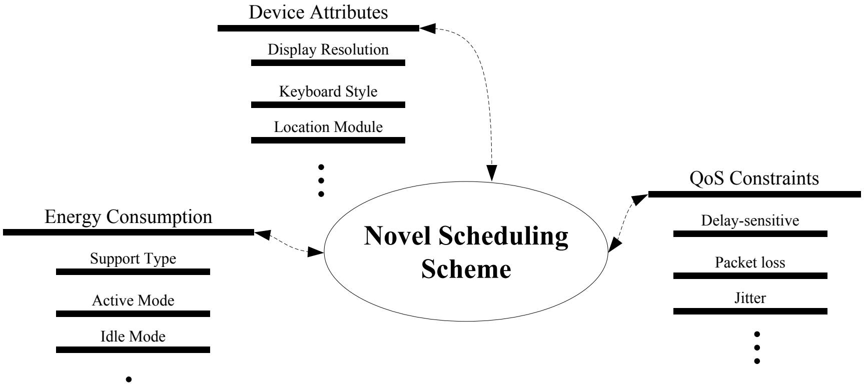


Figure 2. Novel Scheduling Scheme based on Device Attributes, Energy Consumption and QoS Constraints

This paper proposes a novel Scheduling Scheme that makes use of the information about the device attributes, energy consumption of the mobile device, and the QoS constraints as illustrated in Figure 2. Therefore, the resource allocation is done based on a utility function defined as in equation (1). The PRB are allocated to the users with the highest utility.

$$U^{i,j}(t) = [u_r^{i,j}(t)]^{w_r} * [u_e^{i,j}(t)]^{w_e} * [u_{plr}^{i,j}(t)]^{w_{plr}} \qquad (1)$$

where U is the overall utility for stream j of UE i at current scheduling instant t. And $u_r^{i,j}$, $u_e^{i,j}$ and $u_{plr}^{i,j}$ are the utility functions defined for the display resolution of the end-user device, energy consumption of the end-user device and packet loss ratio for UE i, stream j at instant t, respectively. Additionally w_r, w_e, and w_{plr} are the weights for the three criteria, and their sum is 1. It has been shown in [10] that the received bandwidth can be mapped to the user satisfaction for multimedia streaming applications by making use of utility functions.

2.2.1 Display Resolution Utility

In general, the video stream should be played on a display with an adequate resolution in order to ensure a good experience for the mobile users. As various devices have different characteristics and hence different multimedia stream requirements, in this article, we take into account the device resolution when deciding on the device priority. For example, if the device resolution is high, the scheduler will give a high priority, and then the multimedia server will select a high quality level for the multimedia stream. According to the classification in [11], we define the display resolution utility based on different resolutions range as illustrated in Table 1.

Table 1. Utilities of Display Resolutions

	Level 1	Level 2	Level 3	Level 4	Level 5
Resolution	≥1024×768	(1024×768, 768×480]	(768×480, 480×360]	(480×360, 320×240]	<320×240
$u_r^{i,j}(t)$	1	0.75	0.5	0.25	0
	Excellent	Good	Acceptable	Poor	Unacceptable

2.2.2 Energy Utility

Generally, smaller energy consumption ratios are more preferable. Therefore, the energy consumption utility is defined as below:

$$u_e^{i,j}(t) = \begin{cases} 1 & , \quad e_{i,j} \le e_{\min} \\ \dfrac{e_{\max} - e_{i,j}}{e_{\max} - e_{\min}} & , \quad e_{\min} < e_{i,j} < e_{\max} \\ 0 & , \quad otherwise \end{cases} \qquad (2)$$

where $e_{\max}$ is the maximum energy consumption ratio and $e_{\min}$ is the minimum energy consumption ratio among the UEs. And the estimated energy consumption ratio $e_{i,j}$ for UE i for data flow j can be described as in the energy model introduced in [12].

2.2.3 QoS Utility

Packet Loss Ratio is considered for the QoS control in the proposed scheduling scheme. The QoS utility is based on packet loss ratio and is defined as below:

$$u_{plr}^{i,j}(t) = \begin{cases} 1 & , ave\_PLR > target\_PLR \\ ave\_PLR \big/ target\_PLR & , otherwise \end{cases} \qquad (3)$$

where ave_PLR is the average packet loss ratio during a specific time window and $target\_PLR$ is the packet loss ratio tolerance of the video applications.

2.3 Utility-based Prioritization Procedure

The proposed utility-based scheduling scheme is distributed and consists of server-side, mobile-device-side and eNodeB-side. The mobile-device-side is mainly responsibility for collecting the device attributes information, energy consumption rate and QoS responses. This control information is then periodically sent back to eNodeB-side. Moreover, the Server-side integrates a Quality-oriented Adaptation Scheme (QOAS) [8] which ensures the proper transmission of the multimedia streams. The server stores different quality levels of the pre-recorded multimedia streams, from lowest to highest. Based on the feedback received from eNodeB, QOAS adjusts the data rate dynamically. The core function of the proposed scheme is deployed in the eNodeB side. It is located between the MAC layer and PHY layer according to the OSI levels. In this context, a flow diagram of the utility-based prioritization algorithm is defined in Figure 3. After the transmission services start, the coming data flows are queuing in scheduling buffer, and the scheduler is aggregating information of the mobile-device-side, such as display resolutions, energy consumption rates. Once the scheduling buffer is not empty, the resource allocation scheme takes into account the packet loss ratios of data flows and the information of the corresponding mobile devices on receiver-side. And then it computes the overall utilities. Consequently, the data flow with the highest utility is allocated and ready for transmission.

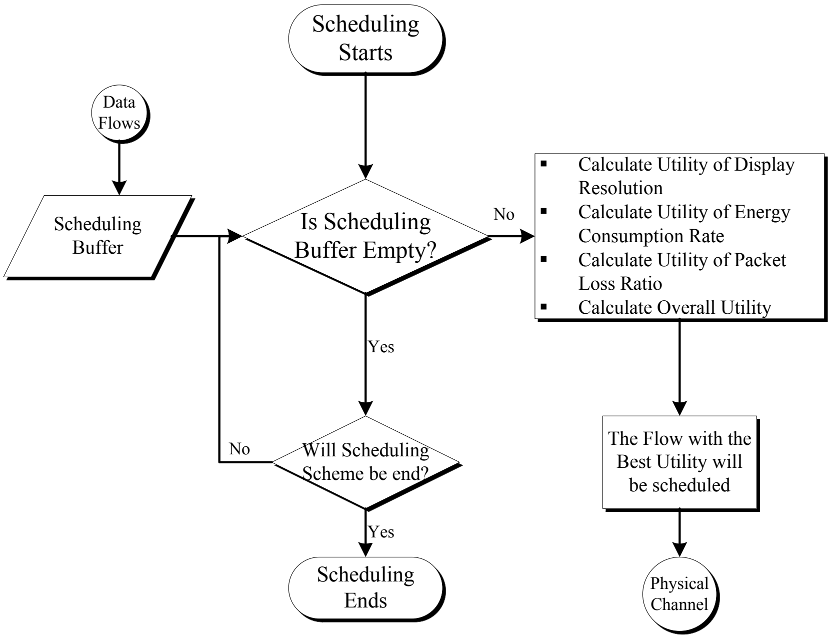

Scheduling Starts
Data Flows
Scheduling Buffer
Is Scheduling Buffer Empty?
No
Calculate Utility of Display Resolution
Calculate Utility of Energy Consumption Rate
Calculate Utility of Packet Loss Ratio
Calculate Overall Utility
Yes
No
Will Scheduling Scheme be end?
The Flow with the Best Utility will be scheduled
Yes
Scheduling Ends
Physical Channel

Figure 3. Utility-based Scheduling Procedure

3 Simulation Environment

Table 2. Simulation Parameters

Parameter	Value	Parameter	Value
Number of eNodeB	1	**Modulation Scheme**	QPSK, 16QAM, 64QAM
Number of UEs	5,10,15,20,25,30,35,40,45,50	**Transmission Power**	20 dBm
Topology	Single Cell	**Transmission Mode**	SISO
User Location	Random Distribution	**Antenna Model**	Isotropic Antenna Model
Cell Radius	1000 meters	**Path Loss Model**	Friis Propagation Model
Carrier Frequency	2.0 GHz	**UE Speed**	3 km/h
Downlink Bandwidth	10 MHz	**Traffic Model**	H.264, Best effort flows, CBR
Number of RBs	50	**TTI**	1 millisecond
Cyclic Prefix	7 Symbols	t_w	10 TTIs

Table 2 lists the simulation parameters used in order to create the simulation environment and validate the. It is assumed that the CQI reporting is error free and the equal downlink transmitting power is allocated to each Physical Resource Block. LTE-Sim [13] is used as the simulation platform, and the parameters of the simulator configuration are listed in Table2. The simulation scenario involves a QOAS server, one eNodeB and several different types of UEs. The goal of these tests is to evaluate the performance of the proposed scheduler compared with the Proportional Fair (PF) Scheduler and Multiclass Modified Largest Weighted Delay First (M-LWDF) scheduler [3], in terms of system throughput or cell throughput, cell packet loss ratio and QoS metrics of videos. The Proportional-Fair Scheduler and M-LWDF are briefly described as in the following equations:

$$scheduling\_metric_{PF}(t,\varphi) = {r_i(t,\varphi)} \Big/ {R_{ave}(t,\varphi)} \tag{4}$$

$$scheduling\_metric_{M-LWDF}(t,\varphi) = \gamma_i \cdot W_i(t) \cdot r_i(t,\varphi) \tag{5}$$

where $r_i(t,\varphi)$ represents the instantaneous rate on PRB φ for user i at time t , and γ_i is a constant whose value is adjusted to different delay requirements of different data flows. $W_i(t)$ is the head of line packet delay of user i at the scheduling instant t and R_{ave} is the moving average of the throughput over a transmission window size t_w . The metrics given by equation (4) and (5) decide the resource allocation priority. For example, the data flow for a user with the highest metric value is given higher priority in resource allocation.

4 Testing and Performance Analysis

For the testing purpose we assume that the QOAS server stores a number of three pre-recorded multimedia quality levels: low quality (e.g., 128kbps), medium quality (e.g., 242kbps), and high quality (e.g., 440kbps). Based on the device characteristics we classify the UEs in three different classes as presented in Table 3. Each class has a corresponding requirement of the multimedia quality level.

Using the simulation parameters listed in Table 2, we conducted a set of different simulation scenarios in which we vary the number of UEs from 5 to 50 which are distributed randomly in a single cell of 1000 meters radius. Each experiment was run three times with different random seeds, which can help to generate the more accurate results with random traffic patterns and path loss distribution. The same simulation conditions were kept when analyzing each of the three schedulers.

Table 3. UE Classification

Type of UE	Class 1	Class 2	Class 3
Display Resolution	480×360	768×480	1024×768
Energy Capacity	5920W	7770W	48000W
Target Packet Loss Ratio	2%	1%	0.1%
Video Trace	128kbps	242kbps	440kbps
CBR traffic	1Mbps	1Mbps	1Mbps
Best effort traffic	Infinite buffer		

Figure 4 illustrates the system packet loss ratio for the video traffic only, with the increase number of users and for each of the scheduling mechanisms. The results are the average values of the three simulation runs. It can be seen that the proposed DE-UPS mechanism outperforms the other two schemes like PF and M-LWDF. For example when having 50 users competing for resources, DE-UPS will reduce the packet loss by 51.8% when compared to M-LWDF.

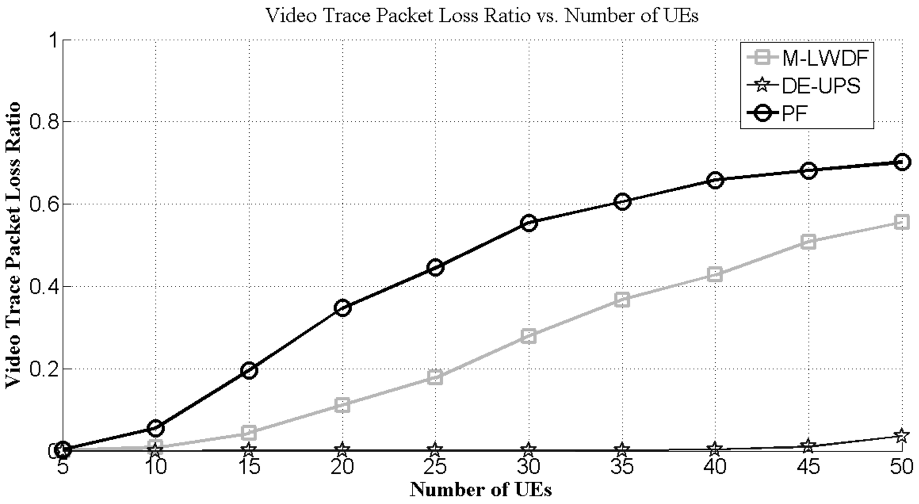


Figure 4. Video Trace Packet Loss Ratio vs. Number of UEs

Figure 5 illustrates the cell throughput of the video trace application under different numbers of UEs. The slops of the curves of PF and M-LWDF become decreasing when the number of UEs is larger than 15. However, the curve of DE-UPS trends to be gentle while the number of UEs is over 45 and the cell throughput could be increasing after 50 UEs. The results show that UPS outperforms PF and M-LWDF in terms of cell throughput. For example when having 50 users competing for resources, DE-UPS will increase the cell throughput by 164.4% when compared to M-LWDF. Additionally, we show a detailed performance comparison of DE-UPS, PF and M-LWDF in Table IV. The experiment with 50 UEs is taken into account for the analysis. And the numbers of different types of UE are generated randomly. Average throughput, energy consumption and packet loss ratio are computed for the different classes of UEs.

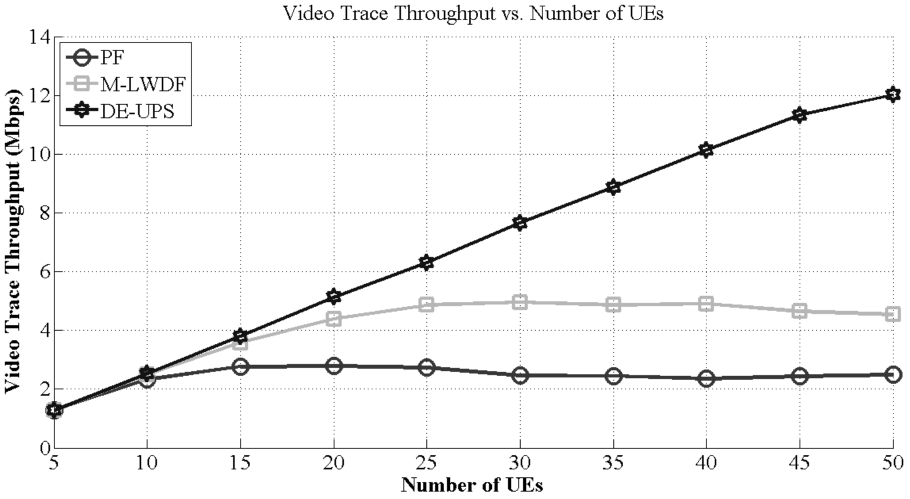


Figure 5. Video Trace Throughput vs. Number of UEs

Table 4 lists the average throughput including the video traffic, CBR traffic and infinite buffer traffic, the average packet loss ratios, and the average estimated energy consumption for each class of UE. The results show that DE-UPS provides a higher priority for the UEs appertaining to the class with high multimedia quality requirements. Thus, because of the device-oriented nature of the proposed solution the system throughput scheduled by DE-UPS is higher than the others. Moreover, the system packet loss ratio of DE-UPS is lower than the others under the same heavy load conditions.

Table 4. Analysis of Comparison Results from Simulation with 50 UEs

Type of UE		Class 1	Class 2	Class 3
Number of UEs		15	23	12
PF	Avg. Throughput (Mbps)	**16.38**	**16.51**	**16.64**
	Ave. Energy Consumption (mW/s)	0.28	0.43	0.14
	Avg. Packet Loss Ratio	0.97169	0.97173	0.97180
M-LWDF	Avg. Throughput (Mbps)	**20.84**	**20.97**	**21.10**
	Avg. Energy Consumption (mW/s)	0.35	0.54	0.18
	Avg. Packet Loss Ratio	0.95911	0.95914	0.95922
DE-UPS	Avg. Throughput (Mbps)	**21.50**	**21.62**	**21.75**
	Avg. Energy Consumption (mW/s)	0.36	0.56	0.18
	Avg. Packet Loss Ratio	0.95758	0.95743	0.95721

5 Conclusion and the Future Works

In this paper we proposed a novel downlink scheduling mechanism for LTE systems when performing video streaming services. The proposed solution is based on a utility function which combines the device characteristics (e.g display resolution) and the energy consumption rate of the mobile device over the transmission channel. With respect to the simulation results and performance analysis, the proposed algorithm allocates a higher number of UE served with acceptable quality in a single cell when compared with the other existing solutions, such as M-LWDF and PF. Future work will consider the fairness between different service types.

Acknowledgements

This work was supported in part by Chinese Scholarship Council, China with Lero-The Irish Software Engineering Research Centre, Ireland.

References

[1] Knopp R. and Humblet P.A., *"Multiple-accessing over frequency-selective fading channels," Personal, Indoor and Mobile Radio Communications*, 1995. PIMRC'95. Wireless: Merging onto the Information Superhighway., Sixth IEEE International Symposium on, vol.3, no., pp.1326, 27-29 Sept. 1995.

[2] Sun S., Yu Q., Yu Meng W., Li C., "A configurable dual-mode algorithm on delay-aware low-computation scheduling and resource allocation in LTE downlink", *Wireless Communications and Networking Conference (WCNC), 2012 IEEE,* vol., no., pp.1444-1449, 1-4 Apr. 2012.

[3] Ramli H.A.M., Basukala R., Sandrasegaran K., Patachaianand R., "Performance of Well Known Packet Scheduling Algorithms in the Downlink 3GPP LTE system", *Communications (MICC), 2009 IEEE 9th Malaysia International Conference on*, vol., no., pp.815-820, 15-17 Dec. 2009.

[4] Comsa S. I., Zhang S., Aydin M., Kuonen P., and Wagen J., "A Novel Dynamic Q-Learning-Based Scheduler Technique for LTE-Advanced Technologies Using Neural Networks", The *37th IEEE Conference on Local Computer Networks (LCN),* Oct. 2012.

[5] Trestian R., Ormond O., and Muntean G-M., "Signal Strength-based Adaptive Multimedia Delivery Mechanism", *The 34th IEEE Conference on Local Computer Networks (LCN),* October 2009.

[6] Qiu X., Liu H., Li D., Zhang S., Ghosal D., and Mukherjee B., "Optimizing HTTP-based Adaptive Video Streaming for wireless access networks," *in 3rd IEEE International Conference on Broadband Network and Multimedia Technology, (IC-BNMT),* 2010.

[7] Chattopadhyay D., Sinha A., Chattopadhyay T., and Pal A., “Adaptive rate control for H.264 based video conferencing over a low bandwidth wired and wireless channel,” *in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, (BMSB),* 2009.

[8] Muntean G.M., Perry P., & Murphy L., “Objective and subjective evaluation of QOAS video streaming over broadband networks,” *Network and Service Management, IEEE Transactions on*, vol.2, no.1, pp.19-28, Nov. 2005

[9] 3GPP, Tech. Specif. Group Radio Access Network - *Physical Channels and Modulation*, 3GPP TS 36.211, Dec. 2009.

[10] Trestian R., Moldovan A., Muntean C-H., Ormond O., and Muntean G., “Quality Utility modelling for multimedia applications for Android Mobile devices,” *in IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB),* 2012, pp. 1–6.

[11] Yuan Z., Venkataraman H., and Muntean G.-M.,"iPAS:An User Perceived Quality-based Intelligent Prioritized Adaptive Scheme for IPTV in Wireless Home Networks", *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB),* Shanghai, China, Mar. 2010, pp.1-6.

[12] Mahmud K., Inoue M., Murakami H., Hasegawa, M., and Morikawa H., “Measurement and usage of power consumption parameters of wireless interfaces in energy-aware multi-service mobile terminals,” *Personal, Indoor and Mobile Radio Communications, 2004. PIMRC* 2004. 15th IEEE International Symposium on, vol.2, no., pp. 1090- 1094 Vol.2, 5-8 Sept. 2004.

[13] Piro G., Grieco L. A., Boggia G., Capozzi F., and Camarda P., “*Simulating LTE cellular systems: An open-source framework*”, *Vehicular Technology, IEEE Transactions on*, 60(2), 498-513, 2012.

MANDL - A **M**ediation **A**lgorithm for **N**etwork Load Balancing Utilising **D**irected **L**earning

Pankaj Goyal [12], Enda Fallon [2], Sidath Handurukande[1], Sheila Fallon[2], Yuansong Qiao[2]

[1] Ericsson Software Research, Athlone, Ireland
pankaj.goyal@ericsson.com, sidath.handurukande@ericsson.com

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

Abstract

Wireless LAN (WLAN) was originally designed to provide coverage in specific “hot spot” areas. The advent of heterogeneous networking has enabled WLAN to play a significant role as a constituent part of a wider IP access network infrastructure. For mobile networks approaches such as cell breathing have traditionally been used to balance network load. Heterogeneous networking requires a re-evaluation of such mobile network specific approaches. In this paper we propose MANDL – A Mediation Algorithm for Network Load balancing Utilising Directed Learning. MANDL is a directed feed forward neural network which uses enforced network handover to balance the load between a set of candidate IP access networks.

Keywords: Network Load Balancing, Wireless Networking, Machine Learning, Directed Learning

1 Introduction

The significant recent consumer interest in mobile applications for IP enabled mobile devices, coupled with enhanced mobile and wireless networking technologies have made heterogeneous networking a reality. The introduction of cellular technologies such as Third Generation networks (3G) together with its variants HSDPA, HSUPA and HSPA+ have enhanced the underlying capability of traditional telecommunications networks. In the area of wireless networks the introduction of 802.11ac and 802.16e has enhanced the range and flexibility of wireless networks.

In conjunction with the increased capability in underlying networks, smart end user devices have attracted significant consumer interest. Many smart phone and computer devices are now designed to support Wi-Fi, Bluetooth, and mobile interfaces. Moreover, many laptops are shipped with standard support for Ethernet, Wi-Fi, WiMAX, Bluetooth, Wireless Infrared, 3G, etc. Underpinned by technical advances in underlying networks and coupled with increased consumer demand for smart devices have resulted in a necessity to support seamless migration between heterogeneous networks.

Wireless LAN (WLAN) was originally designed to provide coverage in specific “hot spot” areas. The advent of heterogeneous networking has enabled WLAN to play a significant role as a constituent part of a wider IP access network infrastructure. The significant number of WLAN installations providing high capacity low cost network access makes it a candidate network for end users. In a heterogeneous IP network infrastructure it becomes necessary to adapt tradition mobile network load balancing approaches to utilise the potential of the wireless infrastructure.

In this paper we propose MANDL – A Mediation Algorithm for Network Load balancing Utilising Directed Learning. MANDL is a directed feed forward neural network which uses enforced network handover to balance the load between a set of candidate IP access networks. MANDL input consists of a selection of normalised performance metrics. Each candidate access network in the configuration is represented by an output neuron. When the stimulation of a neuron exceeds a user defined activation threshold, the neuron “fires” producing a positive output. Positive binary output triggers switchover to the path specified by the inputs. MANDL aims to maximize throughput for all Mobile Nodes (MNs) utilising the network.

The paper is organised as follows; related work is presented in Section 2. The architecture of the MANDL algorithm is presented in Section 3. Simulation-based evaluation is presented in Section 4. Conclusions are discussed in Section 5.

2 Related Work

2.1 Supervised Learning Approaches

The first work on Artificial Neural Networks (ANN) was presented by McCulloch and Pitts in 1943 [1]. 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.

2.1.1 Hebbian Learning

The McCulloch-Pitts network had a fixed set of weights and it was Hebb [2] 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:

$$w_{ij} = \frac{1}{p}\sum_{k=1}^{p} x_i^k x_j^k$$

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. Classification of inputs was introduced by the perceptron model in [3]. The perceptron is a type of artificial neural network invented in 1957 by Frank Rosenblatt. As a linear classifier, it is the simplest kind of feed forward neural network. The perceptron maps its input x (a real-valued vector) to an output value $f(x)$ (a single binary value). The operation of the perceptron can be described as follows:

$$f(x) = \begin{cases} 1 & if\ w.x + b > 0 \\ 0 & Otherwise \end{cases}$$

Where w is a vector of real-valued weights, $w.x$ is the weighted sum of inputs, and b is the 'bias', a constant term that does not depend on any input value. The value of $f(x)$ (0 or 1) is used to classify x as either a positive or a negative instance, in the case of a binary classification problem. If b is negative, the weighted combination of inputs must produce a positive value greater than $|\ b\ |$ in order to push the classifier neuron over the 0 threshold. The bias alters the position of the decision boundary. The introduction of back propagation enabled the training of synaptic weights based on a desired output. Back propagation involved 2 steps: propagation and weight adjustment. Propagation involved the presentation of inputs to the ANN in order to generate output and the comparison of actual and desired output in order to generate a delta value. The weight adjustment stage involves the multiplication of each synaptic weight by a ratio of the delta value. The ratio determines the rate of learning. If the rate of learning is too small, optimization can be centred on local maxima. If the learning rate is too large, the ANN may never reach a trained optimal value.

2.1.2 Supervised Neural Networks

Figure 1 illustrates a supervised learning ANN. Values x_0, x_1, x_2,.... x_n are provided as input to the neuron. The neuron has 2 modes of operation; training and trained. In trained mode, the neuron applies synaptic weights w_{k0}, w_{k1},.... w_{kn} which enhance or degrade the input values. These weighted values are summed and an activation function φ(.) is applied. φ(.) determines whether the neuron should "fire", producing an output y_k which classifies the input pattern.

In supervised learning, the ANN will have an offline training phase in which neural outputs are compared against a training set. Alterations are made to the synaptic weights to limit the error in classification between the output y_k and the training set d_k. When the ANN correctly classifies the input pattern, the ANN operates in trained mode.

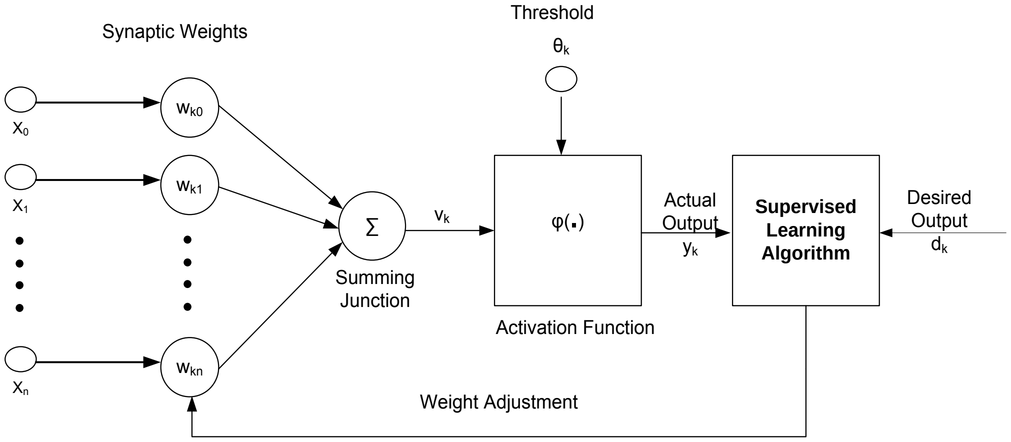


Figure 1: A Supervised Learning Neural Network

There are a number of common activation functions used by feed forward neural networks. A step function is similar to the function used by the original Perceptron. The output is a certain value, e.g. 1, if the input sum is above a certain threshold and 0, if the input sum is below a certain threshold. These kinds of step activation functions are useful for binary classification schemes, when an input pattern has to be classified into one of two groups.

Recent work on ANN enabled load balancing for heterogeneous networks is presented in [4]-[8]. In [4][5][8], the authors propose to improve radio resource utilization for heterogeneous radio systems by selecting the best AP for each terminal and switching the connections dynamically using vertical handover. In order to achieve their aim the authors utilise a McCulloch-Pitts feed forward neural network. As the ANN is a feed forward approach, it has similarities to our MANDL algorithm. When neuron stimulation exceeds a user defined threshold, it fires, producing a binary output indicating that handover should occur to the identified AP. What differentiates the approach from our work is the learning mechanism as their focus is towards load balancing.

In [8] the author's main objective is to maximize average throughput for all users. Alternative objectives of the work focused on minimization of power consumption and communication cost. In [5] the work is extended by introducing a Hopfield neural network type. Hopfield neural networks have similarities to McCulloch-Pitts/Hebbian neural networks however they introduce (a) symmetric weightings between all neurons (b) fully interconnected neurons (each neuron is connected to all other neurons). If a neuron is sufficiently stimulated, it produces a binary output, otherwise the neuron is passive. When the *Energy* and combined output of the network are minimized, the network is considered trained. Hopfield networks have applicability to network load balancing whereby the neurons and interconnections in the network reflect the relationships between MN and APs. By minimizing the Energy of the network, it is possible to minimize e.g. cost or power consumption. Alternatively, applying an inverse, minimizing the Energy of the network can optimize throughput. In [4], the authors utilise their approach to consider both load balancing and end user QoS satisfaction. Unlike our approach which considers QoS enhancement of a specific end user, the approach outlined in [4] focused on optimizing average QoS for all end users on the network. While the authors highlight the relationship between increased throughput and increased QoS, the work concentrates on limiting the differential between required and available throughput for each terminal.

In [6] the authors propose a self-adaptive K value selection scheme for optimizing authentication load balancing in large-scale 802.16 systems. The mechanism considers the cost of authentication, not only at the server end, but also at the client end. The scheme proposes to minimize the overall authentication cost.

In [7] the authors utilise a radial basis function ANN to classify the Uplink Received Signal Strength Indicator (UL RSSI) during high capacity stadium events. A model is derived to assist planning engineers to determine a suitable amount of sectors needed for Distributed Antenna Systems (DAS). The authors utilise experimentally recorded RSSI from the 2010 World Cup as the training set. The number of users is the sole performance metric input to the RBF. Using the experimentally recorded data as the target, the RBF was trained to accurately reflect the relationship between number of users and RSSI.

2.2 Network Load Balancing

Load balancing is the process of dividing and distributing workload between many servers so that more workload can be shared. Load balancing has been mostly used in computer systems for load sharing, it can however also be applied in telecommunications. The aim of resource allocation based NLB is to allocate the resources to the most congested network area. With load distribution, the goal is to use handover to direct the traffic to this congested area.

Load balancing can be implemented in a distributed or centralized way. The centralized approach reduces signalling but is also sensitive to node failure [9]. The distributed approach on the other hand, is simple and robust but requires a greater amount of signalling. It cannot optimize the system in the same way as the centralized approach. When applied to WiMAX, the most appropriate approach is dynamic load balancing implemented in a distributed manner.

Load distribution with handovers can be undertaken in a number of ways. One common approach is cell breathing. The load balancing is implemented by adjusting the transmission levels of the APs pilot signal (shrinking the cell) according to the traffic level. This results in a situation where the MNs at the edge of the cell are forced to implement rescues handovers.

In [10], a new load balancing approach similar to cell breathing in cellular network is proposed. Cell breathing is implemented by controlling the size of WLAN cells, for example, the Access Point's (AP) coverage with the method of changing the transmission power of the AP beacon message dynamically. In [11], a distributed algorithm is proposed in which the 802.11 APs tune their cell size according to their own load and the neighbourhood load.

Cell breathing happens automatically in CDMA systems as the number of MNs increases. This approach can also be used in the other wireless networks, such as 802.11, and Mobile WiMAX. However, there are some disadvantages [12]. A major issue is that a BS would have less control on where and when an MN would conduct the handover. This would limit the possibility to guarantee system wide QoS. In a worst case scenario, the MN will be forced to initiate a rescue handover and might not have a BS in range resulting in communication failure.

[14] proposes a cooperative cellular architecture through which load balancing is achieved by exploiting the broadcast nature of radio waves and using minimum overhead. In this scheme, neither a routing algorithm nor an additional frequency band is required. Both analytical and simulation results are used to evaluate the effectiveness of the proposed architecture. At the same time, the results illustrate that node's selfish behaviour does not significantly affect the performance of the system when only mobile terminals are used as cooperative nodes.

Another NLB approach is traffic load based MN initiated handovers. In this approach the load balancing logic resides in the MN. Such an approach is already used in some WLAN terminals which may choose the least congested AP based on measurements made of the candidate APs. MN initiated load balancing handovers can be easily conducted in Mobile WiMAX based on the available resource information broadcasted in the MOB-NBR ADV message. A handover-based traffic management scheme which can effectively deal with hotspot cells in next-generation cellular networks is proposed in [13]. In the proposed scheme, the traffic load situations of the target cell can be recognized before handover execution.

The above investigations do not take the MN's velocity and direction into consideration when operating NLB. Not considering these factors will result in more handover actions for the MN. In our approach, we use the MIIS to determine the MN's velocity and direction. Handover is directed by the MN's direction as well as the neighbourhood networks load. In this way, we can reduce the occurrence of handovers as well as improving the utilization of network resource.

3 A Mediation Algorithm for Network Load Balancing Utilising Directed Learning

This work proposes MANDL, a directed learning feed forward neural network, which probes network characteristics in order to maximise throughput for all MNs within the network. MANDL operates in either training or trained mode. As an unsupervised algorithm, MANDL does not have an offline training phase. MANDL uses initial end user synaptic weights to determine if handover is required. Following each route cycle, the throughput is calculated and synaptic weights are adjusted. MANDL is trained when (a) the training process does not update synaptic weights (b) synaptic weight updates have no effect on throughput. MANDL ensures that synaptic weights remain relevant to changing network conditions by applying an accuracyThreashold. Throughput is measured for every route cycle and if the accuracyThreashold is exceeded, training is reinitiated.

The MANDL model consists of X_0, X_1,... X_n neuron inputs corresponding to the selected performance metric. Each neuron is a linear threshold gate producing a binary output for path switchover. O_y is defined as follows:

$$V_y = \sum_{i=0}^{N} X_i \, W_i \qquad (1)$$

$$O_y = \begin{cases} 1 \;\; if \; V_y \geq \theta \\ 0 \; if \; V_y < \theta \end{cases} \qquad (2)$$

Figure 2 illustrates the configuration of the MPDLA model. W_{ij} are synaptic weights for each performance metric related to AP_j. V_{yAPj} is the sum of weighted inputs. θ_y is a user configured activation threshold. If the maximum stimulation of all neurons $Max(V_y)$ exceeds the activation threshold θ_y, path handover occurs to the AP with $Max(V_y)$.

The aim is to maximize throughput per route cycle, therefore the rate of change,c, of a linear regression line for historic throughput is calculated as follows:

$$c = \frac{\sum(x-x')(y-y')}{\sum(x-x')^2} \qquad (3)$$

Using c, the rate by which alterations to synaptic weights affect throughput can be determined. A positive c indicates that synaptic weight alterations have a beneficial effect on throughput. A negative c indicates that synaptic weight alterations have a detrimental effect on throughput.

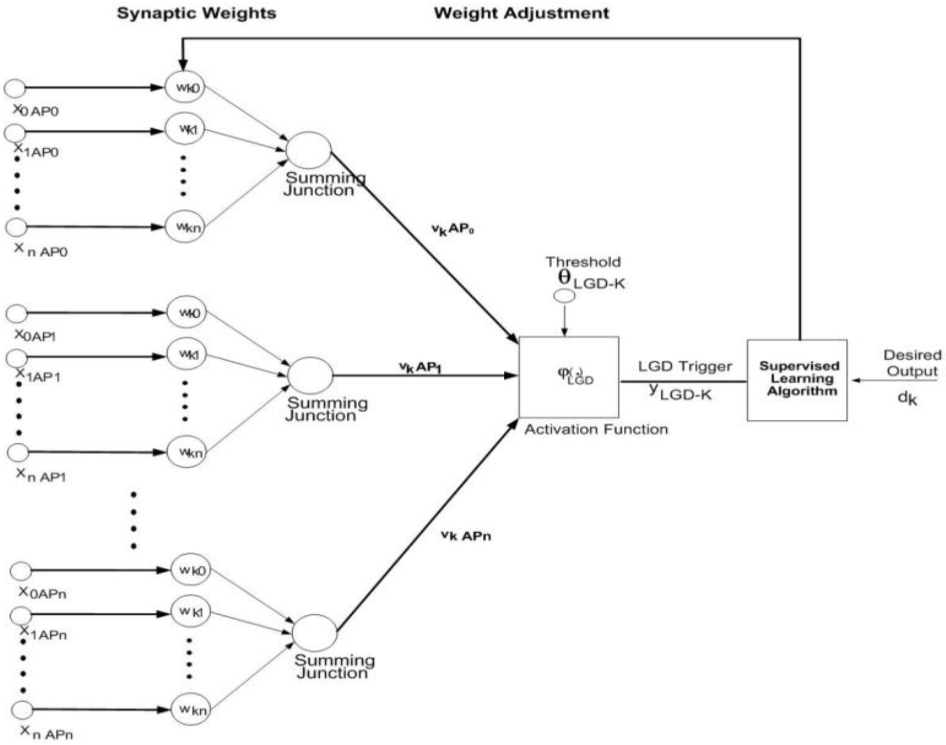


Figure 2: MANDL Algorithm Architecture

In order to control the rate of learning, a user configurable learning rate constant r is defined. 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. The error correction ΔW is defined as the product of c and r.

$$\Delta W = c * r \tag{4}$$

4 SIMULATION-BASED EVALUATION

In this section, we analyse the network performance of MANDL in 2 scenarios; a baseline test is created consisting of 2 APs with 4 MNs, in the second scenarion 15 traffic generating model nodes are connected to the first AP

For analysis, Network Simulation 2 (NS2) is used. Nodes are numbered from zero. 0 and 1 is assigned to two access points AP(0) and AP(1). Numbers 2 to 5 are allocated to the mobile nodes MN(2), MN(3), MN(4) and MN(5). Each AP has a transmit power of 0.051622777W, transmit antenna gain of 1, receive antenna gain of 1 and an antenna height of 1.5m. This provides an outdoor signal range of approximate 716m. The parameters CSThresh(link detection) and RXThresh(link utilisation) were set to -107dBm. MAC technology is 802.11g with frequency 2.472GHz(channel 13). UDP traffic is generated from Mobile nodes using CBR (2 Mbps). A UDP Sink is connected to each Access Points to recieve traffic. Positioning of nodes is linear. Access Points are at (0,0) and (600,0). Mobile Nodes are at (10,0), (280,0), (400,0), (580,0). No Adhoc Routing Protocol is used. Figure 3 illustrates the network configuration. We utilize Throughput, Loss and RSSI as input performance metrics to MANDL. In following section, we outline the results for both scenarios.

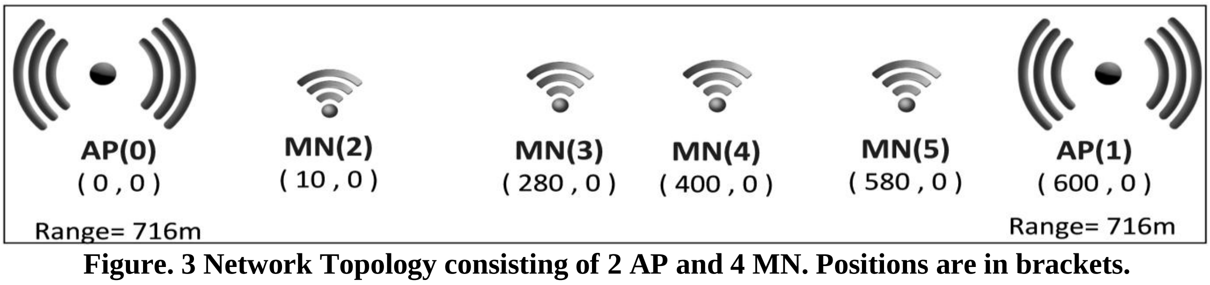


Figure. 3 Network Topology consisting of 2 AP and 4 MN. Positions are in brackets.

4.1 Optimum Weight Learning And Configuration For 4 Node Infrastructure Network.

Initially no node is connected to any AP. We connect each MN to both APs one by one. Table 1 illustrates (a) the performance metrics when each MN is connected to each AP (b) normalised performance metrics (c) the initial weight assignment (d) the output (activation value Vy) when the weights are applied to the input metrics for each AP and MN.

		Inputs			Normalized			Weights			
AP	**MN**	**Throughput**	**Loss**	**RSSI**	**Throughput**	**Loss**	**RSSI**	**W1**	**W2**	**W3**	**Output**
0	**2**	1953	0	-33.0	0.977	1.000	0.870	*0.5000*	*0.4800*	*0.1000*	**1.0554**
1	**2**	1953	0	-103.8	0.977	1.000	0.162	*0.5000*	*0.4800*	*0.1000*	**0.9845**
0	**3**	1953	0	-90.8	0.977	1.000	0.292	*0.5000*	*0.4800*	*0.1000*	**0.9975**
1	**3**	1953	0	-93.2	0.977	1.000	0.268	*0.5000*	*0.4800*	*0.1000*	**0.9952**
0	**4**	1953	0	-97.0	0.977	1.000	0.230	*0.5000*	*0.4800*	*0.1000*	**0.9913**
1	**4**	1953	0	-85.0	0.977	1.000	0.350	*0.5000*	*0.4800*	*0.1000*	**1.0033**
0	**5**	1953	0	-103.5	0.977	1.000	0.165	*0.5000*	*0.4800*	*0.1000*	**0.9848**
1	**5**	1953	0	-45.0	0.977	1.000	0.750	*0.5000*	*0.4800*	*0.1000*	**1.0433**

Table 1: Initial Performance Metrics of Each MN Connected to Each AP, Normalized metrics, Weights and Output

Each performance metric is scaled and normalized. The scale for throughput is 0-2000 Kb, for Loss its 0-100 % and for RSSI its 20-120dBm. Initial weights are applied on these normalized values to get the output (W1*Throughput+W2*Loss+W3*RSSI). The selection of an AP is based on this output value. A performance threshold of 1 is applied. If both the output for both AP exceeds the threshold the AP with the highest output value is chossen as the Point of Attachment (PoA). The following calculation illustrates the selection of an AP for MN(2) .

$$\left\{\begin{array}{l}\text{Output1} = 0.977\text{x}0.5000 + 1.000\text{x}0.4800 + 0.870\text{x}0.1000 = \mathbf{1.0554.}\\ \text{Output2} = 0.977\text{x}0.5000 + 1.000\text{x}0.4800 + 0.162\text{x}0.1000 = \mathbf{0.9845.}\end{array}\right\} \text{AP(0) is selected}$$

So after this iteration, the topology is as follows.
MN(2) → AP(0) MN(3) → not connected MN(4) → AP(1) MN(5) → AP(1)

This topology yields a total throughput for all MNs, 5860 Kb. Now the value of weights will be modified based on the slope of upto 5 previous iterations. New weights are calculated and applied and outputs are compared leading to topology changes. Throughput of next iteration will lead to slope changes and weight adjustment. Training this network repeatedly will give us the optimum value of weights for maximum throughput. Table 2 illustrates the learning process.

w1	**w2**	**w3**	**Thresh.**	**Iteration**	**Throughput**	**Slope**	**Error Correction**	**Learning Rate**	**Topo.**
				0	0			0.0001	
0.5000	0.4800	0.1000	1	1	5860	5859.96	0.58600	0.0001	***0- -1-1***
1.0860	1.0660	0.6860	1	2	7813.3	3906.64	0.39066	0.0001	***0-0-1-1***
1.4767	1.4567	1.0767	1	3	7813.3	2539.32	0.25393	0.0001	*0-0-1-1*
1.6840	1.5205	0.8523	1	4	7813.3	1757.99	0.17580	0.0001	*0-0-1-1*
1.8598	1.6963	1.0281	1	5	7813.3	390.66	0.03907	0.0001	*0-0-1-1*
1.8988	1.7354	1.0671	1	6	7813.3	0	0.00000	0.0001	*0-0-1-1*
1.8988	1.7354	1.0671	1	7	7813.3	0	0.00000	0.0001	*0-0-1-1*

Table 2: MANDL Based Learning Process

After the first learning cycle, c = 5859.96, r = 0.0001 and ΔW = 0.586 (using eq. 4) results in a weight assignment w1=1.08, w2=1.066, w3=0.68 for the second traversal of the route. Every fourth cycle a random weight adjustment is applied to each weight in order to avoid the potential of learning being concentrated on local maxima. Second learning cycle results in throughput of 7813.3 and weight adjustment of 0.39066. Last column shows topology after every iteration. After few iteration, we get constant weights with zero adjustment. Figure 4 shows variation of weights with learning cycles.

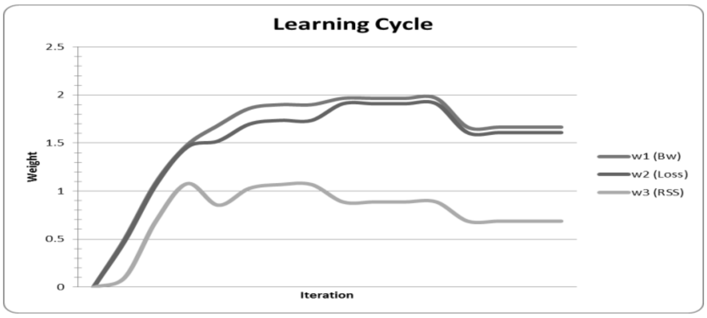


Figure 4. Graph between learning cycles and weights.

4.2 Optimum Weight Learning And Configuration For 4 Node Infrastructure Network With Bruteforce On One Access Point

In this section we reevaluate the optimal weight configuration following the introduction of 15 traffic generating nodes connected to AP(0). Figure 5 illustrates the configuration.

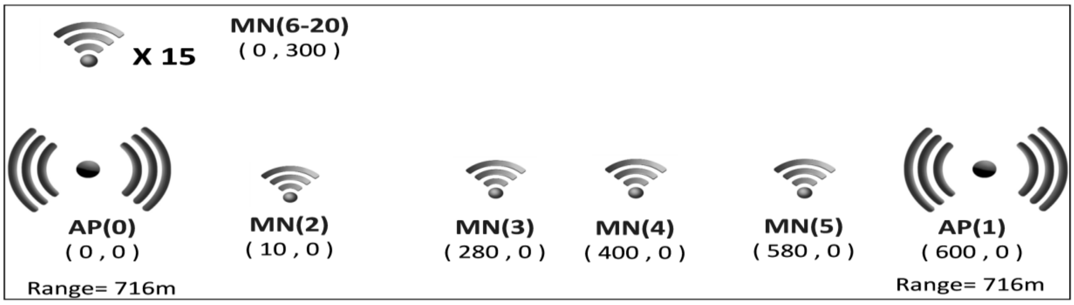


Figure 5. Network Topology consisting of 2 AP and 4 MN with Bruteforce.

Table 3 illustrates that MN(3), MN(4) and MN(5) experience a significant degradation in performance when connected to AP(0) when the 15 traffic generating nodes are added. MN(2) is close to AP(0) and does not experience any degradation. MN(3) experiences a drop in throughput from 1953 to 1258 packets. MN(2) and MN(3) experience similar throughput fall off. As a result of the performance degradation the output (activation value) for MN(3) connected to AP(0) decreases from .9975 to .6529. Similar degradations are experienced from MN(4) and MN(5), however these MNs are already connected to AP(1) so the degradation is not relevant.

		Inputs			Normalized			Weights			
AP	MN	Throughput	Loss	RSSI	Throughput	Loss	RSSI	W1	W2	W3	Output
0	**2**	1953	0	-33.0	0.977	1.000	0.870	*0.5000*	*0.4800*	*0.1000*	**1.0554**
1	**2**	1217	37.7	-103.8	0.608	0.623	0.162	*0.5000*	*0.4800*	*0.1000*	**0.6195**
0	**3**	1258	35.6	-90.8	0.629	0.644	0.292	*0.5000*	*0.4800*	*0.1000*	**0.6529**
1	**3**	1953	0	-93.2	0.977	1.000	0.268	*0.5000*	*0.4800*	*0.1000*	**0.9952**
0	**4**	1251	36	-97.0	0.625	0.640	0.230	*0.5000*	*0.4800*	*0.1000*	**0.6431**
1	**4**	1953	0	-85.0	0.977	1.000	0.350	*0.5000*	*0.4800*	*0.1000*	**1.0033**
0	**5**	1183	39.4	-103.5	0.592	0.606	0.165	*0.5000*	*0.4800*	*0.1000*	**0.603**
1	**5**	1953	0	-45.0	0.977	1.000	0.750	*0.5000*	*0.4800*	*0.1000*	**1.0433**

Table 3: Performance metrics of each MN connected to each AP, Normalized metrics, Weights and Output (15 Traffic Generating Nodes Connected to AP(0))

Table 4 illustrates the MANDL based learning process when 15 traffic generating nodes are connected to AP(0). It illustrates that following the initial training cycle described in Table 3 the MNs are connected to the following APs; MN(2)-AP(0), MN(3)-Not Connected, MN(4)-AP(1), MN(5)-AP(1). Table 4 illustrates that between learning iteration 1 and 2 the weights W1, W2 and W3 increase from 0.5000, 0.4800, 0.1000 respectively to 1.0859, 1.0659, 0.6859. The addition of the 15 traffic generating nodes degraded throughput and loss but not RSS. The increase in weight allocation for throughput and loss to 1.0859, 1.0659 causes a large deviation in the performance evaluation of AP(0) and AP(1). This is particularly pertinent for MN(3) which can be potentially connected to either AP.

After iteration1, our topology becomes and remains:
MN(2) → AP(0)
MN(3) → AP(1)
MN(4) → AP(1)
MN(5) → AP(1)

w1	w2	w3	Thresh.	Iteration	Throughput	Slope	Error Correction	Learning Rate	Topo.
				0	0			0.0001	
0.5000	0.4800	0.1000	1	1	5859	5859	0.58590	0.0001	**0- -1-1**
1.0859	1.0659	0.6859	1	2	7810.2	3905.08	0.39051	0.0001	**0-1-1-1**
1.4764	1.4564	1.0764	1	3	7810.2	2538.16	0.25382	0.0001	**0-1-1-1**
1.3057	1.4431	1.1195	1	4	7810.2	1757.15	0.17571	0.0001	**0-1-1-1**
1.4814	1.6188	1.2952	1	5	7810.2	390.231	0.03902	0.0001	**0-1-1-1**
1.5204	1.6578	1.3342	1	6	7810.2	0	0.00000	0.0001	**0-1-1-1**
1.5204	1.6578	1.3342	1	7	7810.2	0	0.00000	0.0001	**0-1-1-1**

Table 4: MANDL Based Learning Process with 15 Traffic Generating Nodes

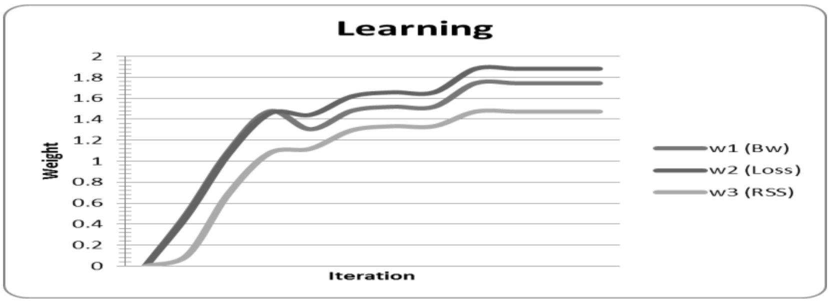


Figure 6. Graph between learning cycles and weights

Fig 6 illustrates a sudden increase in wieight allocation following the addition of the nodes. Banwidth and Loss have a higher relative weight compared to RSS. This is as a result of increased network congestion on AP(0) rather than a degradation in received RSS.

5 CONCLUSION

In this work we propose MANDL – A Mediation Algorithm for Network Load balancing Utilising Directed Learning. MANDL is a directed feed forward neural network which uses enforced network handover to balance the load between a set of candidate IP access networks. MANDL input consists of a selection of normalised performance metrics. Each candidate access network in the configuration is represented by an output neuron. When the stimulation of a neuron exceeds a user defined activation threshold, the neuron “fires” producing a positive output. Positive binary output triggers path switchover to the path specified by the inputs. MANDL aims to maximize throughput for all MNs utilising the network. We evaluate MANDL in a homogeneous wireless networking configuration. We degrade the performance of one of the candidate access networks and illustrate that the learning approach implemented by MANDL quickly reevaluates network conditions and reconfigures MN PoA in order to optimise total network throughput.

Future work will consider heterogeneous networking configurations in which both wireless and mobile access networks are utilised as candidate PoA.

References

[1] W. McCulloch, W Pitts *A logical calculus of the ideas immanent in nervous activity*

[2] D. Hebb "*The Organization of Behavior*", published by Wiley

[3] F. Rosenblatt "*A comparison of several perceptron models*."

[4] M. Hasegawa, K. Ishizu, H. Murakami, H. Harada *Experimental evaluation of distributed radio resource optimization algorithm based on the neural networks for Cognitive Wireless Cloud* Proceedings of IEEE PIMRC Workshops, pp. 32-37 2010

[5] M. Hasegawa et.al *Design and Implementation of A Distributed Radio Resource Usage Optimization Algorithm for Heterogeneous Wireless Networks* Proceedings of VTC Fall, pp. 1-7 2009

[6] F. Yuang, N. Xiong, J.H. Park *A self-adaptive K selection mechanism for re-authentication load balancing in large-scale systems* Journal of Super Computing 15th July 2011

[7] S. Roets, P. Reddy, P. Govender *UL RSSI as a design consideration for distributed antenna systems, using a radial basis function model for UL RSSI* Proceedings of the 5th international conference on Network optimization Inoc 2011

[8] M. Hasegawa, N. Ha, G. Miyamoto, Y. Murata *User-Centric Optimum Radio Access Selection in Heterogeneous Wireless Networks Based on Neural Network Dynamics*, in Proceedings Wireless Communications and Networking Conference, 2008. WCNC 2008. IEEE , vol., no., pp.2747-2752, March 31 2008-April 3 2008

[9] T. Casey, N. Veselinovic, and R. Jantti, *Base Station Controlled Load Balancing with Handovers in Mobile WiMAX*, PIMRC 2008. IEEE 19th International Symposium on Personal, Indoor and Mobile Radio Communications, pp.1-5, December 2008.

[10] Bejerano, Y., Seung-Jae Han, *Cell Breathing Techniques for Load Balancing in Wireless LANs*, IEEE Transactions on Mobile Computing, June 2009, vol. 8, pp. 735 – 749

[11] Garcia, E., Vidal, R., Paradells, J., *Cooperative load balancing in IEEE 802.11 networks with cell breathing*, Computers and Communications, 2008. ISCC 2008. IEEE Symposium on, Marrakech, July 2008, pp. 1133 – 1140.

[12] S. H. Lee and Y. Han, *A Novel Inter-FA Handover Scheme for Load Balancing in IEEE 802.16e System* IEEE 65th Vehicular Technology Conference, 2007, pp. 763 - 767, April 2007.

[13] D. Kim, M. Sawhney, and H. Yoon, *An effective traffic management scheme using adaptive handover time in next-generation cellular networks* International Journal of Network Management, Volume 17, Issue 2, pp. 139 - 154, March 2007.

[14] Ghaboosi, N. Jamalipour, A., *A Cooperative Cellular Architecture with Emphasis on Traffic Load Balancing*, 2010 IEEE Wireless Communications and Networking Conference (WCNC), April 2010, pp 1

Reputation-based Network Selection Solution in Heterogeneous Wireless Network Environments

Ting Bi, Ramona Trestian and Gabriel-Miro Muntean

Performance Engineering Laboratory, Dublin City University, Ireland
E-mail: ting.bi2@mail.dcu.ie, ramona@eeng.dcu.ie and munteang@eeng.dcu.ie

Abstract

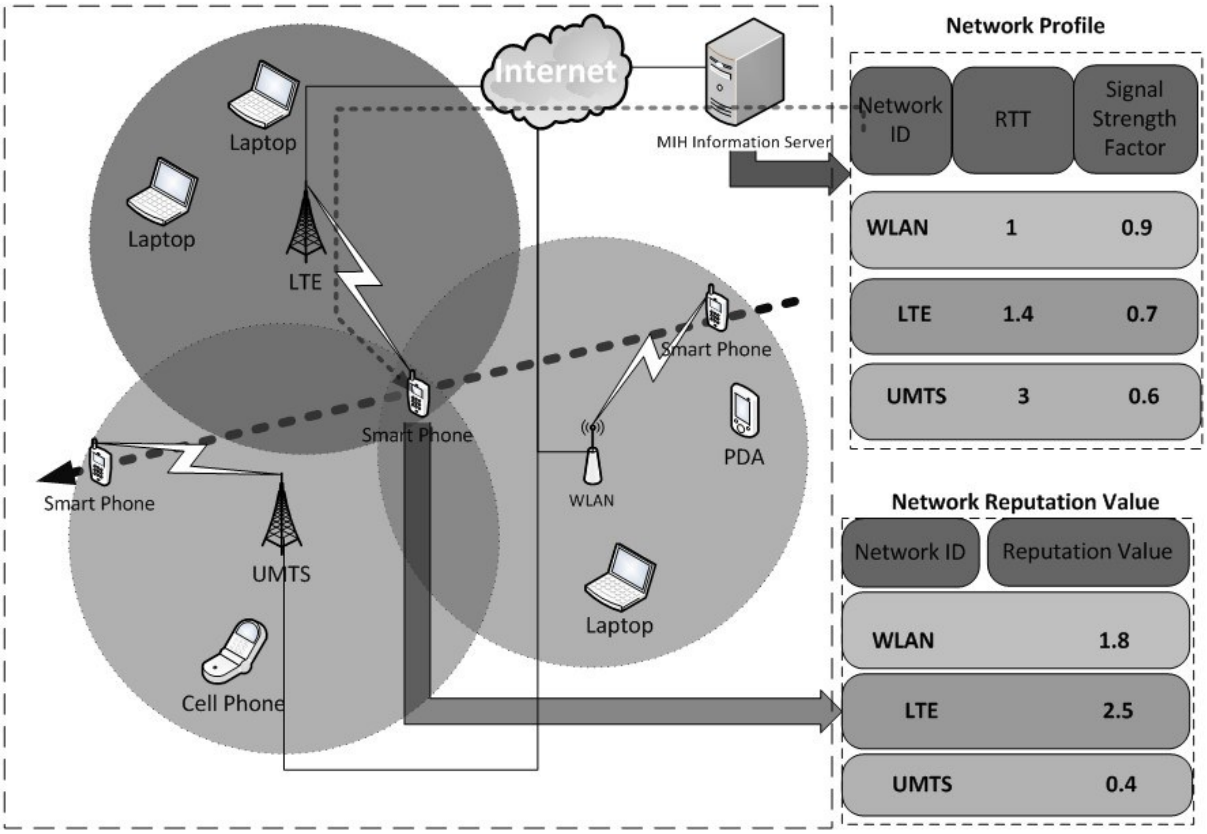
The significant developments in terms of both mobile computing device (e.g., smartphones, tablets, laptops, etc.) and the wireless communication technologies (e.g., LTE, LTE-Advanced, WiMAX, etc.), lead towards a converged heterogeneous wireless environment. In this context, the user will be facing the problem of selecting from a number of Radio Access Networks that differ in technology, coverage, pricing scheme, bandwidth, latency, etc. In order to provide high quality of service (QoS) to the user in this heterogeneous wireless environment, a network selection solution is required that will efficiently facilitate the vertical handover between different wireless access networks in a seamless manner. In this paper, we propose a reputation-based network selection solution which aims to select the best value network for the user. We propose a network profiling algorithm that used to compute the reputation of each of the available networks based on the joint collaboration of the users within the network. The network with the best reputation value is recommended for selection and handover.

Keywords: Reputation Mechanism, Heterogeneous Wireless Network, Network Selection, vertical handover

1 Introduction

Due to the rapid evolution of cellular and wireless networks together with the advances in technologies and the rapid adoption of mobile computing devices led towards a multi-technology multi-terminal multi-application multi-user heterogeneous wireless environment representing the next generation of wireless networks. In this context the "Always best Connected" vision emphasis the scenario of variety of radio access technologies that work together in order to provide global wireless infrastructure in which the en-users will benefit from an optimum service delivery via the most suitable available wireless network that satisfies their interests. However, supporting such a connectivity goal is very difficult, mostly due to system complexity and diversity of technologies.

In terms of wireless technologies, wireless networks are grouped into three major categories: Wireless Local Area networks (WLAN), Wireless Wide Area networks (WWAN) and Wireless Personal Area Networks (WPAN). WLAN networks are mostly represented by the IEEE 802.11 family (i.e. including the well-known 802.11a/g/b/n and the recent IEEE 802.11ac) and they offer high data delivery rates, but they have limited transmission range. WWAN networks provide coverage over extremely large areas, best known for their Global System for Mobile Communications (GSM) and Universal Mobile Telecommunications System (UMTS) technologies, the latest Long-Term-Evolution (LTE) protocol provides support for higher data rates which could reach 3 Gbps downlink and 1.5 Gbps uplink [1]. WPAN networks are the smallest wireless networks used to connect various peripheral devices centered around an individual person's workspace. The two kinds of wireless technologies used for WPAN are Bluetooth and Infrared Data Association.



Network ID	RTT	Signal Strength Factor
WLAN	1	0.9
LTE	1.4	0.7
UMTS	3	0.6

Network ID	Reputation Value
WLAN	1.8
LTE	2.5
UMTS	0.4

Figure 1. Heterogeneous Wireless Networks Scenario & Information Server Sample

The "optimally connected anywhere, anytime" vision was introduced by ITU in Recommendation ITU-R M.1645 [2] in June 2003 and relies on different radio access networks connected via flexible core networks. The aim is to provide seamless, transparent and QoS-enabled connectivity to the user by taking into account the limitations of the underlying wireless access technology and user preferences.

Moreover, the IEEE 802.21 Media Independent Handover (MIH) Working Group [3] (Jan 2009) considers the interoperability aspects between heterogeneous networks, and has developed a standard referred to as IEEE 802.21. This standard enables the optimization of handover between heterogeneous IEEE 802 networks and facilitates handover between IEEE 802 networks and cellular networks by providing a media-independent framework and associated services. However, the new standard does not provide a network selection mechanism itself.

In this context, this paper proposes a novel Reputation-based Network Selection Solution (RNS) to enable the best selection of a candidate wireless network. RNS is based on the joint collaboration of the users within a network and makes use of the IEEE 802.21 MIH standard mechanisms in order to gather performance information about the current wireless network from the users. This performance information is then aggregated and disseminated to other mobile users in the form of a network profile. The network profile is used to make an informed quality-oriented decision when selecting the candidate network for handover in the heterogeneous wireless network environment.

For example, Fig.1 illustrates a scenario of a mobile user roaming through a heterogeneous wireless environment. In this particular case, there are three different access network technologies considered: WLAN, UMTS and LTE. For each of the networks, network profile stored in the MIH information server. A mobile user (MU) using a smartphone device can be located within the coverage area of a WLAN. Following user mobility, MU can face the choice of selecting between three wireless networks. In this context, the MU can send a request to the MIH Information Server which responds with the network profile of the available networks. Based on the response information, MU can generate the reputation value of each network and make a network selection, then will handover to the new network. The network profile for each network is stored in the MIH information server. In Fig.1, those values are pseudo value for numerical analysis, also shows in Table 1.

The remaining of this paper is organized as follows: Section II discusses related works. In Section III, detailed information about the RNS system architecture and the RNS algorithm are described and discussed. Section IV provides the performances results, and, finally, Section V presents the conclusions and future work directions.

2 Related Work

Reputation systems have been studied and deployed to the wireless environment [4], especially in mobile ad-hoc networks, wireless mesh networks, and Internet-based peer-to-peer, being useful in cooperation scenarios and decision making problems. For example, reputation systems are used in order to help peers decide with whom to cooperate or not. Peers with good reputation are favored.

In [5] an enhanced MIH Information Server to accelerate vertical handover procedures in the 802.21 framework is proposed. They reduce the vertical handover latency by eliminating time-consuming channel scanning procedure. Authors in [6] proposed an energy-aware utility-based user-centric network selection strategy in heterogeneous wireless network environments, which is using the Media Independent Handover Function (MIHF) to gather and exchange information. In [7] an enhanced IEEE 802.21 MIH based framework that integrates a Vertical Handover Management Engine (VHME) for vertical handover decision-making based on networks reputation is described. The authors make use of a large set of parameters that map the QoS and QoE to a network reputation value.

Some other papers [8-10] describe reputation-based network selection strategies and vertical handover solutions. In [8], the authors present a reputation based VHO decision rating system by proposing the use of the grey model first order one variable (GM(1,1)). Their proposed solution provides a quick and efficient prediction of reputation score for a target network in the handover decision making progress. The QoS parameters like Bit Error Rate (BER), delay, jitter and bandwidth are used to calculate the reputation value for UMTS, WiMAX and WLAN networks. The proposed solution was evaluated through simulations using the network simulator NS2. The results show that the reputation-based system can provide the mobile node with advance time to make a successful handover and thus experience an overall higher QoS. Zekri et al in [9] propose a VHO management solution combining the use of reputation as a Quality of Experience (QoE) indicator for fast decision-making. This solution collects individual user experience. By users expressing their past experiences, the system aggregated those individual score to give a reputation value for Wi-Fi, WiMAX and UMTS networks. The performance results show that this solution provides better handover latency and throughput than other solutions. Trestian et al. in [10] propose a reputation-based network selection mechanism using game theory. The user-network interaction is modeled as a repeated cooperative game and the reputation of the network is computed based on the user's payoff. The proposed solution is based on individual user experience and the mechanism is integrated into an extended version of the IEEE 802.21 model.

In all these previous related works, multi-user involvement in information gathering or network reputation building and reputation information exchange has not been considered. These are the main contributions of this paper.

3 Architecture

3.1 System architecture

The RNS block-level architecture is shown in Fig. 2. This system architecture consists of two main components: Mobile Nodes (MN) and a MIH Information Server.

In order to perform network selection, the MN needs the list of candidate networks and also their associated quality levels. 802.21 MIH provides a mechanism to support gathering and exchanging of information between various network components, MIH Information Server and MN. Each of the

MIH-enabled entities contains a cross-layer MIHF. This function provides Service Abstraction Points (SAP) acting as an abstract interface between a service provider and a user entity. The higher-layer user entities employ the MIH-SAP to control or monitor the link-layer entity, and the MIHF uses the MIH-LINK-SAP as an interface together with the link layer to translate the data received from the MIH-SAP. The remote MIHF entities use the MIH-NET-SAP to exchange the information with the MIHF.

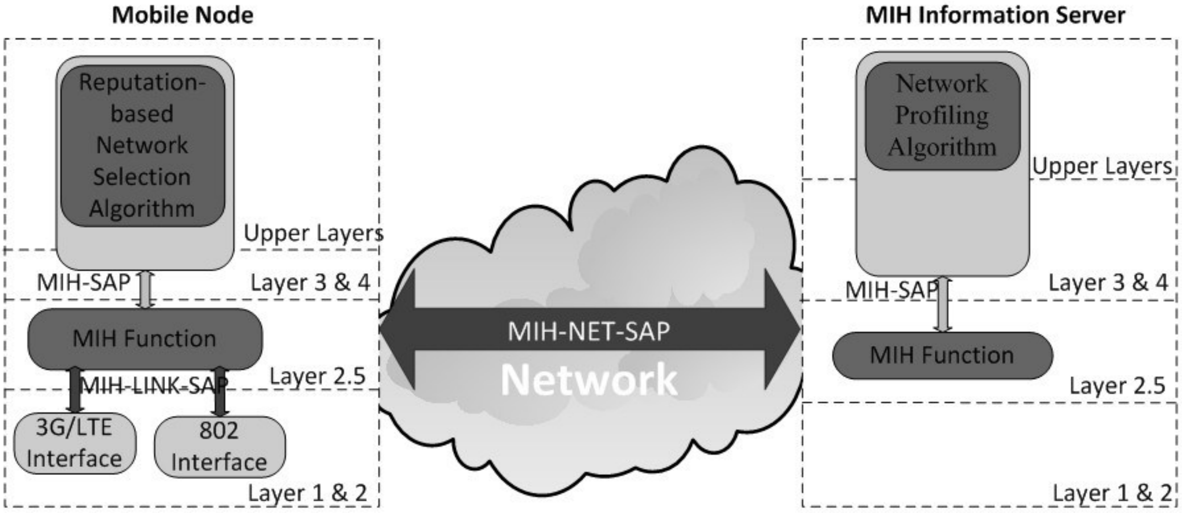


Figure 2. System Architecture

The proposed RNS solution is distributed and consists of a server side component and a client side component. At the server side, the Network Profiling Algorithm (NPA) is based on the joint collaboration of the users within a network. The MIH Information Server gathers the performance information feedback from multiple users within the network and computes the performance factor for that particular network. At the client side, Reputation-based Network Selection Algorithm (RNSA) using the network profile gathered from MIH Information Server computes the network reputation value. The network with the highest reputation value is selected as the target network and the handover process is triggered.

3.2 Network Profiling Algorithm (Server Side)

In order to execute the Network Profiling Algorithm (NPA), the data required includes: current access point (AP) location (X_{ap}, Y_{ap}), MN location (X_{mn}, Y_{mn}), and signal strength of current network i measured at user's current location $Su^{i}_{(Xmn,Ymn)}$.

For a wireless channel model, the theoretical value of the signal strength of current network i at user's location $St^{i}_{(Xmn,Ymn)}$ is calculated based on distance d between AP and MN under COST-231 Hata model[11] by using the signal strength equation described in [5][11][12] and in equation (1):

$$PL(d)_{dB} = 46.3 + 33.9\log_{10}(f) - 13.82\log_{10}(h_b) - a(h_r) + (44.9 - 6.55\log_{10}(h_b))\log_{10}(d) + c_m \quad (1)$$

In equation (1), $PL(d)_{dB}$ is the path loss expressed in dB, f is the carrier frequency, h_b is the antenna height at the AP, and d is the distance between the AP and MN:

$$\mathrm{d} = \sqrt{\left(X_{mn} - X_{ap}\right)^2 + \left(Y_{mn} - Y_{ap}\right)^2} \quad (2)$$

$a(h_r)$ is the MN's antenna height correction factor and h_r is the MN's antenna height. The parameter c_m is a constant with values *3 dB* and *0 dB* for urban and suburban environments, respectively. By using equation (3), the theoretical value of the signal strength of user's location $St^{i}_{(Xmn,Ymn)}$ could be obtained [5]:

$$\mathrm{St}^{i}_{(X_{mn},Y_{mn})} = P_{tdB} - PL(d)_{dB} \quad (3)$$

Where P_{tdB} is the transmit power expressed in dB.

Finally, the signal strength utility value $U^{i}_{SS}$for the user in current network *i* at position *(X_{mn}, Y_{mn})* is computed using $Su^{i}_{(Xmn,Ymn)}$, $St^{i}_{(Xmn,Ymn)}$ and equation (4).

$$\mathrm{U}^{i}_{SS} = \begin{cases} 1 & , \mathrm{Su}^{i}_{(X_{mn},Y_{mn})} \geq \mathrm{St}^{i}_{(X_{mn},Y_{mn})} \\ \frac{\mathrm{Su}^{i}_{(X_{mn},Y_{mn})}}{\mathrm{St}^{i}_{(X_{mn},Y_{mn})}} & , \mathrm{Su}^{i}_{(X_{mn},Y_{mn})} < \mathrm{St}^{i}_{(X_{mn},Y_{mn})} \\ 0 & , Otherwise \end{cases} \quad (4)$$

The user regularly sends performance reports (UPR) which can be described as multi-tuple as in equation (5):

$$\mathrm{UPR} = [\mathrm{MNID}, (X_{mn}, \mathrm{Y}_{mn}), \mathrm{U}^{i}_{SS}, \mathrm{NetworkID}, \mathrm{RTT}](5)$$

Where *MNID* is the ID that identifies the mobile node and *NetworkID* identifies the current network. RTT is the Round-Trip Time between the times t_1 which MN sends the report and the time t_2 which MN receives the response from MIH information server:

$$\mathrm{RTT} = t_2 - t_1 \quad (6)$$

NPA is presented in pseudo code in Algorithm below. It describes how the signal strength factor F^{i}_{SS} and average RTT $\overline{RTT_i}$ for each network *i* can be generated given the utility function value $\mathrm{U}^{i}_{SS}$and RTT received from any reporting node located in that network. NRR_i is the number of performance reports for network *i* received so far.

Algorithm 1: Network Profiling Algorithm

```
If (first report) then
    Initialize $NRR_i=0$
if $(NRR_i = 0)$ then
$F^{i}_{SS} = \mathrm{U}^{i}_{SS}$; $\overline{\mathrm{RTT}}_i = \mathrm{RTT}$;
else
$\mathrm{F}^{i}_{SS} = \frac{\mathrm{F}^{i}_{SS}*\mathrm{NRR}_i+\mathrm{U}^{i}_{SS}}{\mathrm{NRR}_i+1}$;$\overline{\mathrm{RTT}}_i = \frac{\overline{\mathrm{RTT}}_i*\mathrm{NRR}_i+\mathrm{RTT}}{\mathrm{NRR}_i+1}$;
end if
$\mathrm{NRR}_i$++;
```

3.3 Reputation-based Network Selection Algorithm (Client Side)

The RNSA is located at the MN and it based on the network profile report (NPR) sent by the MIH Information Server.

$$\mathrm{NPR} = (\mathrm{NetworkID}, \mathrm{F}^{i}_{SS}, \overline{\mathrm{RTT}}_i, \mathrm{NRR}_i) \quad (7)$$

The reputation value *R_i(X_{mn}, Y_{mn})* for network *i* at user's current location *(X_{mn}, Y_{mn})* can be calculated as in equation (8):

$$R_i^{(X_{mn},\mathrm{Y}_{mn})} = \frac{F^{i}_{SS} \times \mathrm{Su}^{i}_{(X_{mn},Y_{mn})}}{\overline{RTT}_i} \quad (8)$$

RNSA is presented in pseudo code in algorithm below. Once MN receives the response which consist the NPR, the RNSA algorithm is executing to generate the reputation value of available networks. By comparing the reputation value of available networks $R_i^{(X_{mn},\mathrm{Y}_{mn})}$and the current network $R_c^{(X_{mn},\mathrm{Y}_{mn})}$, MN selects the highest one as the target network to execute the handover. After this, MN will send a new UPR to MIH information server to update the network profiles that make the reputation mechanism accurate.

Algorithm 2: Reputation-based Network Selection Algorithm

```
Initial: receives the response from MIH information server, $\mathrm{RTT} = t_2 - t_1$;
if ($\overline{\mathrm{RTT}}_i = 0$) then
    $\overline{\mathrm{RTT}}_i = RTT$ ;
else
    $\overline{\mathrm{RTT}}_i = \frac{\overline{\mathrm{RTT}}_i * \mathrm{NRR}_i + \mathrm{RTT}}{\mathrm{NRR}_i + 1}$ ;
end if
$R_i^{(X_{mn},Y_{mn})} = \frac{F_{SS}^i \times \mathrm{Su}^i_{(X_{mn},Y_{mn})}}{\overline{RTT}_i}$;
If($\exists(R_i^{(X_{mn},Y_{mn})} > R_C^{(X_{mn},Y_{mn})})$) then
    Select the highest one as the target network to execute the handover.
else
    Send new UPR to MIH information server.
```

3.4 Network selection and handover

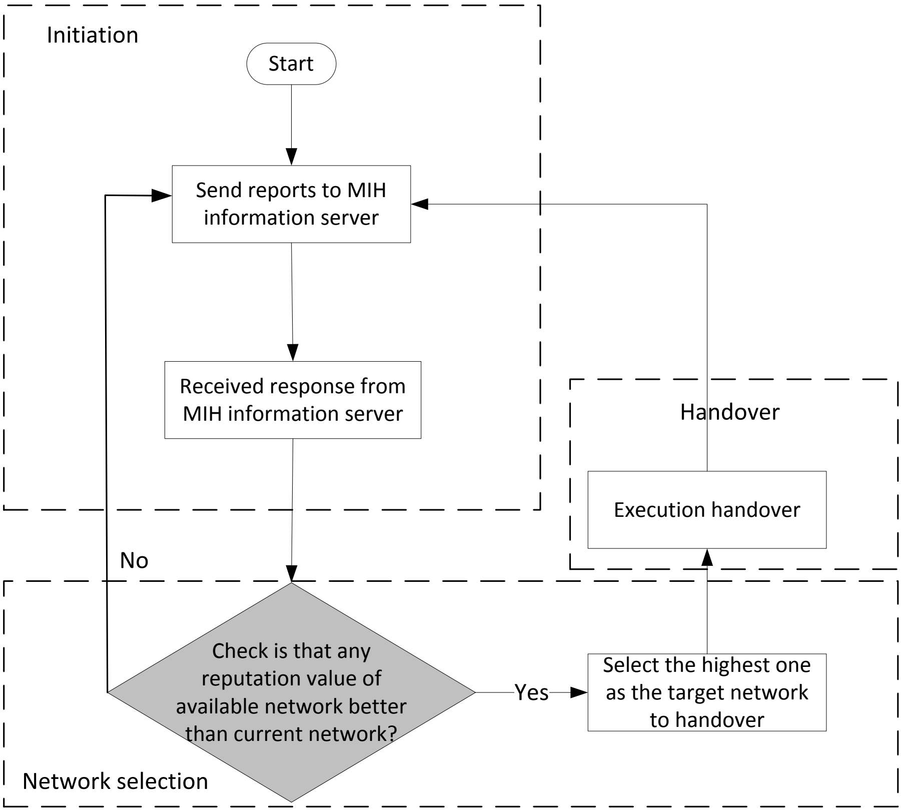

Initiation
Start
Send reports to MIH information server
Received response from MIH information server
Handover
Execution handover
No
Check is that any reputation value of available network better than current network?
Yes
Select the highest one as the target network to handover
Network selection

Figure 3. Network Selection and Handover Mechanism

The proposed solution consists of three main phases: Initiation, Network Selection and Handover Execution as depicted in Fig.3. MN sends an information request and user report to the current serving attachment point when it initiates a connection with the current serving network. The current serving network forwards this information request to the MIH information server. The MIH Information Server receives information requests, and user reports from MNs using MIH-NET-SAP. On receiving any information, MIH Information Server sends it from MIHF to the upper-layers in charge with network selection-related data storage and processing, and immediately responds to MN. The information response extends the 802.21 MIH protocol with one additional field: Network Profile Report. This report contains the list of the candidate networks along with signal strength factor, average RTT and NRR. Based on the network profile report, the MN generates the reputation value for each of the candidate networks. The candidate network with the highest reputation value is selected as the target network. Finally, MN executes handover from current network to target network. Once MN success handover to the target network, MN will send a new UPR to MIH information server contains the RTT value of the previous network and signal strength utility factor of the new network.

4 Numerical Analysis

Table 1. Network Profile for Each Network

	$\overline{\mathrm{RTT}}_i$	F^i_{SS}
UMTS	3	0.6
LTE	1.4	0.7
WLAN	1	0.9

This section describes the simulation scenario and the numerical results and analysis. This scenario considers the case of a typical day in a business professional life which travels from point A (e.g., home) to point E (e.g., office) as illustrated in Fig 4. The mobile user will pass through three different networks: UMTS, LTE and WLAN. On the way to the office the mobile user needs to be always connected to the internet. Thus, the RNSA is enabled in the user's mobile device and the MN will select the best network to handover to in order to support "always best connect" internet service. The values in Table 1 are pseudo value for each network for numerical analysis, and Table 2 is also using the pseudo signal strength level instead of real data. Based on the data provided in Table 1 and Table 2, and by using the equation (8) the reputation values for each network are listed in Table 3.

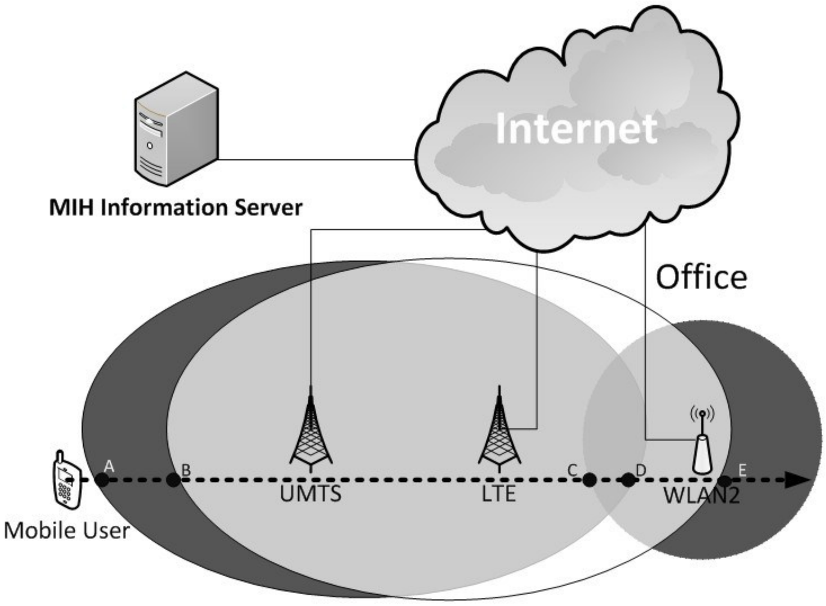


Figure 4. Simulated Scenario

Table 2. Signal Strength level for Each Network

	A	**B**	**C**	**D**	**E**
UMTS	1	**3**	2	1	-
LTE	-	2	**5**	**4**	2
WLAN	-	-	2	**4**	**5**

Table 3. Reputation Value of Each Network

	A	**B**	**C**	**D**	**E**
UMTS	**0.2**	0.6	0.4	0.2	-
LTE	-	**1**	**2.5**	2	1
WLAN	-	-	1.8	**3.6**	**4.5**

From the Table 3, at point B the MN will handover to LTE network, even if the signal strength level of UMTS is better than LTE at this point, by considering the network reputation based on both signal strength and RTT, the reputation value of LTE is better than the one for UMTS. At point C the MN will not handover to WLAN network until the MN moved to the point D. In point D the signal strength level of both LTE and WLAN are the same, but WLAN has a better reputation factor in both signal strength and RTT. Finally, user reaches the destination at point E.

5 Conclusions and Future Work

This paper proposes a reputation-based network selection solution in heterogeneous wireless network environments. Based on the location of MN, signal strength of MN and the RTT of heterogeneous networks, the proposed solution selects the most appropriate network for the user.

Numerical results show that the proposed algorithm achieves a good reputation value in heterogeneous wireless network environments. And the network selection and handover mechanism will support the "always best connected" paradigm.

Further work will use the network simulator NS3 to evaluate the proposed algorithm under various scenarios. Location information from user reports to estimate user route and therefore future user position relative to various networks' coverage areas will be considered as the next step.

Acknowledgement

The authors would like to thank the support of Irish Research Council Enterprise Partnership with Everseen Ltd.

References

[1] LTE-Advanced Rapporteur, 'RP-090939: 3GPP Submission Package for IMT-Advanced', www.3gpp.org, 3GPP TSG RAN, meeting 45, Seville, Spain, September, 2009.

[2] ITU, ―Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-200, Switzerland, ITU-R M.1645, 2003.

[3] IEEE 802.21-2008, Standard for Local and Metropolitan Area Networks-Part 21: Media Independent Handover Services, *IEEE Computer Society*, Jan. 2009.

[4] Buchegger S. et al., "Reputation systems for self-organized networks*, IEEE Technology and Society Magazine,* vol. 27, no. 1, pp. 41 – 47, 2008.

[5] Kim, Y., Pack, S., Kang, C. G., and Park, S. (2011). An enhanced information server for seamless vertical handover in IEEE 802.21 MIH networks. *Computer Networks,* 55(1), 147–158.

[6] Trestian, R., Ormond, O., and Muntean, G.-M. (2010). Power-friendly access network selection strategy for heterogeneous wireless multimedia networks. *2010 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB),* Shanghai, China, 1–5

[7] Zekri, M., Jouaber, B., & Zeghlache, D. (2012). An enhanced media independent handover framework for vertical handover decision making based on networks' reputation. *37th Annual IEEE Conference on Local Computer Networks -- Workshops*, 673–678.

[8] Giacomini, D., & Agarwal, A. (2012). Vertical handover decision making using QoS reputation and GM(1,1) prediction. *2012 IEEE International Conference on Communications (ICC)*, 5655–5659.

[9] Zekri, M., & Pokhrel, J. (2011). Reputation for Vertical Handover decision making. *Communications (APCC), 2011 17th Asia-Pacific Conference on*, (October), 318–323.

[10] Trestian, R., Ormond, O., and Muntean G-M. (2011), "Reputation-based network selection mechanism using game theory," *Physical Communication*, vol. 4, no. 3, pp. 156–171, Sep. 2011.

[11] Trestian, R., Ormond, O. and Muntean, G.-M. (2009). Signal Strength-based Adaptive Multimedia Delivery Mechanism, *the 34th IEEE Conference on Local Computer Networks (LCN),* Zürich, Switzerland, 297–300.

[12] COST ACTION 231, "Digital Mobile Radio Towards Future Generation System," *Technical Report, European Communities*, EUR 18957, 1999.

Session 2

Information Analysis and Management

Gathering Transportation Data by Acoustic Monitoring: A Case Study

Vi Tran-Ngoc-Nha[1], Patrick McDonagh[2], John Murphy[1]
[1] LERO@UCD, School Of Computer Science and Informatics
University College Dublin, Belfield, Dublin 4, Ireland.
vi.tran-ngoc-nha@ucdconnect.ie, j.murphy@ucd.ie
[2] LERO@DCU, School Of Electronic Engineering
Dublin City University, Glasnevin, Dublin 9, Ireland.
patrick.mcdonagh@dcu.ie

Abstract

Acoustic data is a potential source for traffic monitoring due to its low-cost and the ease of deployment. In this paper, a case study of using acoustic monitoring as a source for transportation management purposes is conducted. The results show the feasibility of detecting different traffic conditions by analyzing audio waveforms. An application is also developed to generate a large number of audio samples. The purpose of building this application is to prepare a database for further research work on performing complex and continuous queries on transportation data.

Keywords: Traffic Management, Transportation, Acoustic Monitoring.

1 Introduction

Intelligent Transportation Systems (ITS) handle huge amounts of data from various sources. How to derive timely and precise traffic information from this large amount of data is challenging. Even though there is a massive amount of data, it is still insufficient to provide real-time information [1]. Therefore, transportation systems require not only more data but also more accurate data to supplement existing data sources such as Closed-circuit television (CCTV) video streams, induction loop systems and GPS information.

Acoustic sensor monitoring has several potential advantages. Firstly, it is low-cost and simple to install compared to CCTV or induction loop systems [2]. Secondly, acoustic data can be used alone or to supplement other sources. Finally, since the audio file size is small compared to video, acoustic data can be sent with lower transmission costs over wired or wireless networks.

In this paper, we conduct a case study on how to detect different traffic patterns from acoustic monitoring sources. The result highlights the possibility of classifying acoustic patterns of different traffic conditions from audio waveforms. The recognized patterns are: with a single car passing, with multiple cars passing, background noise without car passing and with congestion. We have also developed an application to generate a large number of acoustic samples by randomly combining these patterns. This is a preparatory step to build a database for our future research on performing continuous queries from multiple data sources in an ITS.

The structure of this paper is as follow. Section 2 provides a review on related work in the domain of acoustic monitoring for traffic management. In Section 3, we present our case study using recorded audio samples to classify the acoustic pattern of different traffic conditions. In Section 4, we introduce an application to generate acoustic traffic data by joining the classified traffic patterns. Section 5 presents the proposed architecture for processing continuous transportation queries. We then conclude the paper and discuss our future works.

2 Related Works

Research on acoustic traffic monitoring has been conducted on various subtopics. One primary research direction in acoustic monitoring is to study the side effects of traffic noise pollution on urban life [3]. Recently, acoustic monitoring has also been considered as a data source for traffic management to detect flow-rate of a street, a vehicle speed and to classify different types of vehicles.

In [4], Averbuch proposed an algorithm to detect the passing of a vehicle by analyzing its acoustic signature against an existing database of recorded and processed acoustic signals. This study applied concepts of pattern matching for vehicle type classification. In [5], the authors use noise emission of various vehicles of different size to classify them into four different categories: light motor vehicle, medium heavy vehicles, heavy vehicles and powered motorbikes. They also find that acceleration/deceleration frequency of motor/ light vehicle is from 50 to 250 Hz.

In [6], the authors focus on vehicle classification and vehicle detection using traffic sound analysis. They found that for an accelerating vehicle, the amplitude of the sound increases steadily, while a moving vehicle with constant speed has a constant amplitude pattern. From this work, a sound database is built from 3 classes of vehicles: two wheeler, three-wheeler and heavy vehicle. The data set collected had 300 sound clips sampled at 5 kHz, 100 of each type with a clip duration varying from 2 to 8 seconds ensuring they had predominant sound of the vehicle. This extraction was done manually by listening to the recordings and viewing the waveform and spectrogram.

In [2], Barbagli presents research on the energy distribution of both passing vehicles and standing vehicles and propose a acoustic wireless sensor network to detect vehicles as well as showing the correlation between energy distribution and the sound map. This creates a prototype with experimental results showing that data regarding vehicle numbers and vehicle speed can be derived from acoustic sensor networks. They validate their work with data from an induction loop system.

The authors of [7] introduce their passive acoustic traffic monitoring system and findings. They find that a microphone array can be used for vehicle counts, or to detect vehicle presence, speed or direction. They also claim that the bandwidth of sound from vehicles is approximately in the range of 300-800Hz.

Not all of the research focuses on detecting vehicles or vehicle type. For example, Sen et al. [8][9] also proposed an idea of using acoustic sensing to detect traffic congestion. Their study is different from ours in two aspects. Firstly, they conduct experimental evaluation in Mumbai, India while our case study is in Dublin and Ho Chi Minh with different types of vehicle and driving behaviors. They consider the amount of vehicular honks as an indicator to calculate vehicle speed and vehicle distribution via their proposed hardware prototype *Doppler*. This system uses two sensors with 20 meters away from each other to estimate vehicle velocity by recording frequency change when a vehicle passes by these sensors. In our case, we detect traffic condition by analyzing the energy level patterns of the recorded audio signals.

3 A Case Study of using Acoustic Monitoring as a Source for Transportation Data

In this section, we will describe how we design and carry out our case study and provide a discussion of the results.

The aim is to find an indication of audio energy for car detection as well as the patterns of different traffic conditions based on the acoustic data sources. In order to do so, we analyse the waveform of recorded audio files. In the future, we want to build an acoustic database based on the discovered patterns.

3.1 Case Study Description

Several video clips with audio were recorded on the N11 dual-carriage-way in Dublin city, for both congested and free-flow conditions. The main vehicles on the road were cars and buses. There were also pedestrians walking and cyclists passing by. Furthermore, we recorded video and audio data of different traffic conditions in Ho Chi Minh city, Vietnam where there is a significant number of motorbikes on the street. According to World Bank data [11], vehicular trips in Vietnam urban area consist of approximately 65% motorbikes, 5% automobiles and 7% buses. The chosen roads were open space and there were also pedestrians and cyclists on the road.

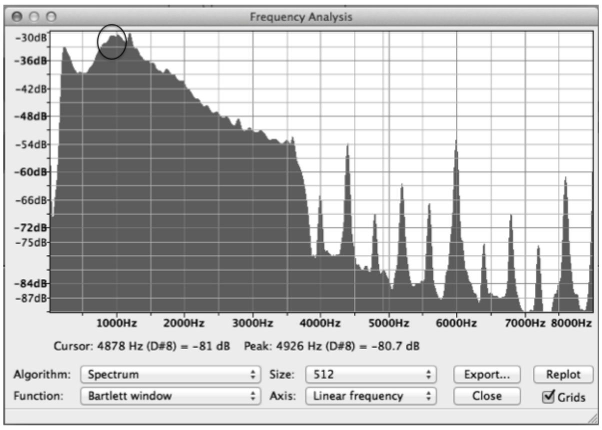
The weather conditions when the clips were taken place were dry and calm. We expect that during wet weather, there might be more background noise in the recorded sound. In our results, the analysis was performed for only the clips recorded under normal weather conditions.

All the clips were captured by the application using the built-in video camera and microphone of a HTC Nexus One mobile phone. We chose Audacity [10] as the tool for audio processing since it is open-source and supports the necessary operations on audio files, such as equalization frequency analysis, etc. At first, the audio clips need to be imported to Audacity. Audacity supports viewing files by spectrogram or waveform diagrams. The waveform shows the amplitude of an audio signal over time, while the spectrogram reflects the density of the frequency components over time. To inspect the dominant frequency of an audio file, the Spectrum Plot feature is used. Audacity also supports effects such as Equalization, High/Low Pass and Bass Boost. With the Equalization effect, we can emphasize a desired frequency for more detailed analysis.

For the purpose of analysing our recorded clips, we first divide them into free-flow and congested categories.

3.2 Free-flow Condition

Initially we performed equalization to emphasize the energy in the range of 300-800 Hz for free-flow conditions. Kuhn et al.[7] states that vehicle sound is from that frequency range. However, there is limited success when we analyze the audio files with frequency filtering in the range of 300Hz-800Hz. Therefore, to find the emphasized frequency, we study the Spectrum Plot of free-flow clips. A sample spectrum of a free-flow clip is shown in Figure 1. The highest peak of energy level is from the range 800Hz-1.25KHz approximately. The explanation for the frequency difference from Kuhn study is that recorded audio is the combination of rolling of tyre noise with vehicle noise [5]. The highlighted energy peak is considered as an implication of the presence of car in our sample. Equalization is performed with the band 800Hz-1.25KHz. Figure 2 is the spectrogram of the free-flow condition before and after equalization. The peak in this diagram indicates a passing car (verified by watching video). The circles in Figure 2 are the gap in filtered frequency where no car is detected. The pattern of multiple cars passing and no cars passing are also identified in Figure 3. They are patterns allowing us to identify free-flowing traffic and similar conditions.



Cursor: 4878 Hz (D#8) = -81 dB Peak: 4926 Hz (D#8) = -80.7 dB

Algorithm: Spectrum
Size: 512
Function: Bartlett window
Axis: Linear frequency
☑ Grids

Figure 1: Sample Spectrum Plot of a Free-flow clip

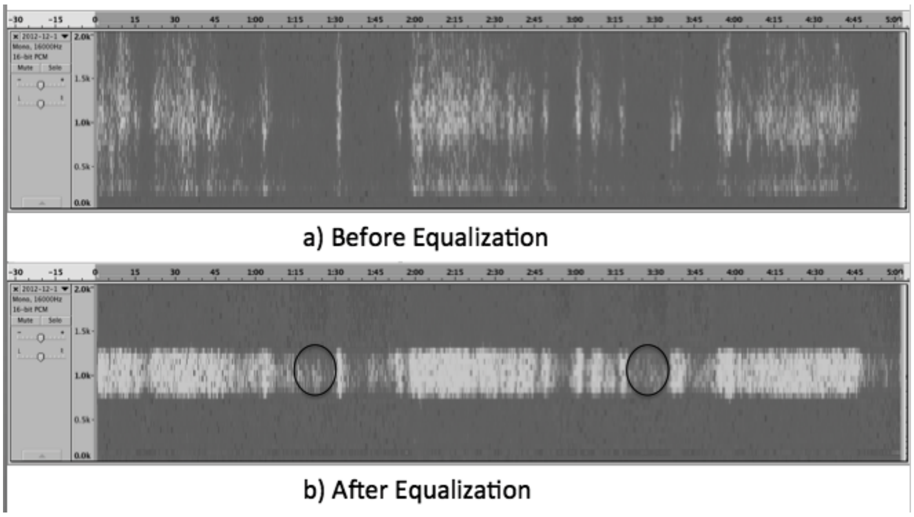


Figure 2: Free-flow spectrogram before (above) and after (below) equalization

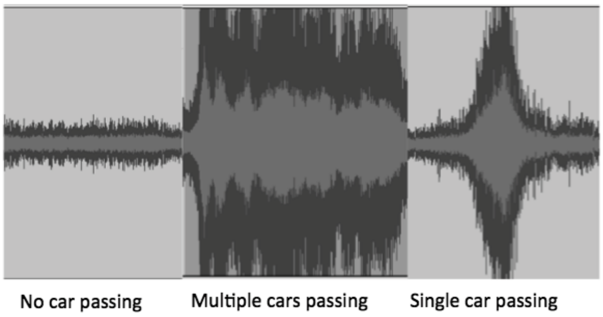


Figure 3: Traffic Waveform Patterns of a) No cars passing b) Multiple cars passing c) Single car passing

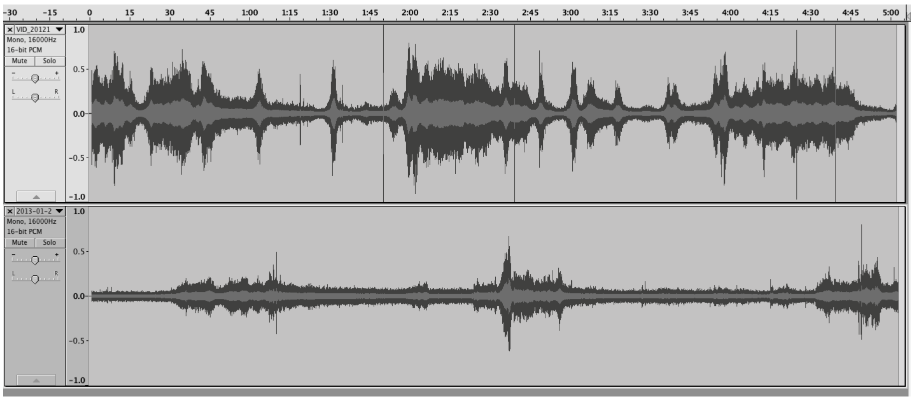
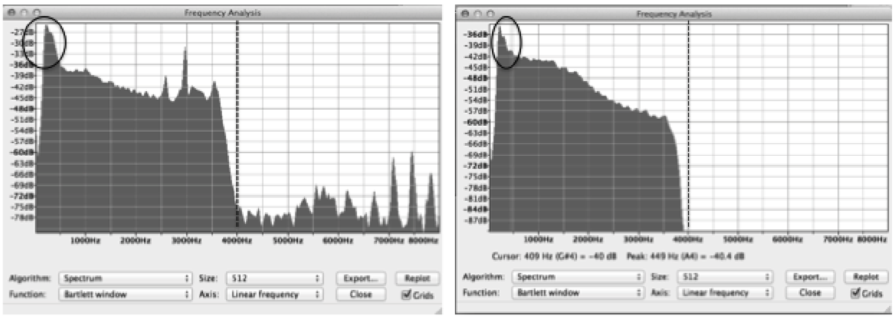


Figure 4: Waveform of both Scenarios a) Free-flow (above) and b) Congestion (below)

3.3 Congested Condition

For congested conditions, we expect that peak-to-average energy level is lower on the displayed waveform diagram. Figure 4 shows the difference of waveform diagrams for both conditions. The Spectrum Plot of congestion clips (both in Ho Chi Minh and Dublin) show prominent energy level in the range of 200Hz-500Hz in Figure 5. Higher frequency noise from the Ho Chi Minh clips is the result of vocal, honk and other noise being reflected with the presence of frequency above 4KHz. However, we performed equalization only on the 200-500Hz dominant range and how the spectrogram changes are shown in Figure 6. The white color in this figure is the presence of energy the frequency range over time.



Algorithm: Spectrum Size: 512
Function: Bartlett window Axis: Linear frequency

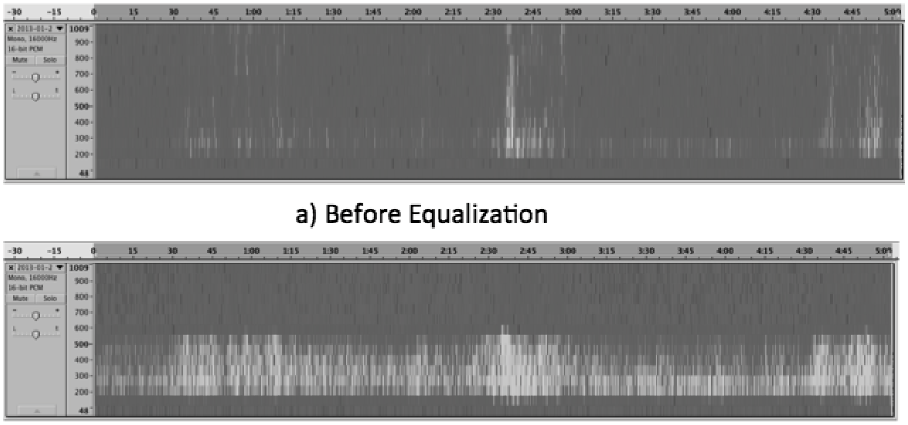
Cursor: 409 Hz (G#4) = -40 dB Peak: 449 Hz (A4) = -40.4 dB
Algorithm: Spectrum Size: 512
Function: Bartlett window Axis: Linear frequency

a) Ho Chi Minh Congestion

b) Dublin Congestion

Figure 5: Sample Spectrum Plot of a) Ho Chi Minh Congestion and b) Dublin Congestion

This white area also has less interruption gap than free-flow spectrogram. This is different than free flow condition and also an indication of traffic congestion.



a) Before Equalization

b) After Equalization

Figure 6: Congestion spectrogram before (above) and after (below) equalization

4 An Application of Traffic Audio Generator

We now have audio samples of different traffic patterns: no cars passing, one car passing, multiple cars passing and traffic congestion. Our plan is to build an acoustic database to feed to a continuous query engine of an ITS. An application is developed to generate a large numbers of traffic acoustic samples.

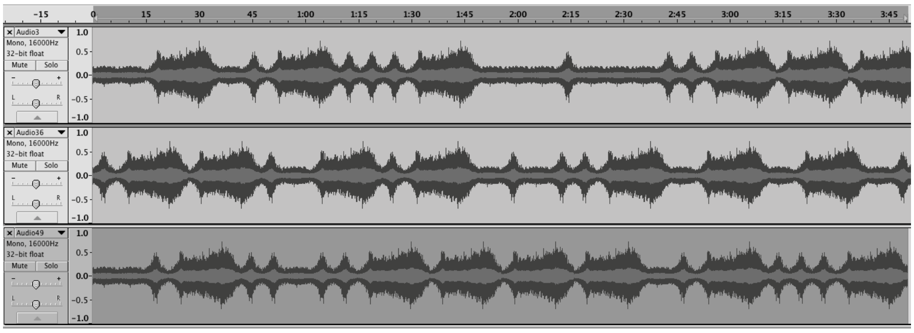


Figure 7: Waveforms of sample generated traffic sound

The program is written in Java and use Java Sound API (javax.sound.sampled) for randomly generating sample traffic audio files. If we wish to generate clips that conform to a certain pattern, we could weight the probability that a certain sample is chosen. This API provides the implementation for the processing of audio data. The main classes used for our application are AudioFileFormat, AudioInputStream, AudioSystem. AudioSystem supports the reading or writing audio files while AudioInputStream is used to process them. The generated files are stored in the Wave format.

Figure 7 is the waveform of sample generated by our application. In order to generate these files, the first step is to identify three patterns and stored them in three audio files. The next step is generate a new clip by combining those identified patterns. A randomly selected pattern is appended to the clip until its duration is longer than 5 minutes and the clip is stored. This process is performed for 50 times to create 50 random clips. However, as stated before, a clip representing a particular pattern could be generated. Figure 8 shows the workflow of this process.

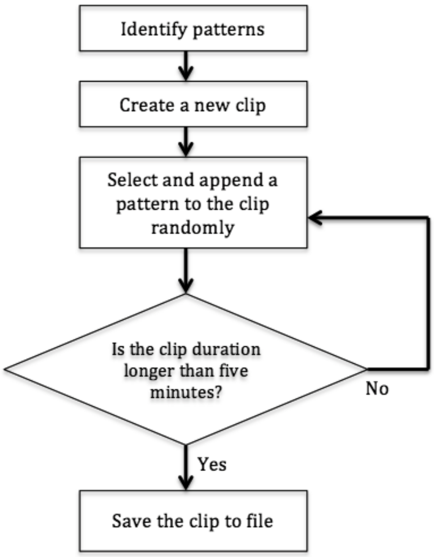

Identify patterns
Create a new clip
Select and append a pattern to the clip randomly
Is the clip duration longer than five minutes?
No
Yes
Save the clip to file

Figure 8: Process Workflow

5 Architecture of Continuous Transportation Queries

Our vision is to process continuous transportation queries. Data for the ITS is continuously updating and fed to the system as data streams. Moreover, transportation information requires real-time processing. Queries for the ITS therefore, must be continuous. Building a catalogue and preparing traffic data from different sources are first steps of this work. We aim to retrieve timely and accurate traffic information by streaming query processing of multiple data sources. Using the architecture to process continuous queries proposed in Babu's seminal paper [12], we propose the architecture for processing transportation queries shown in Figure 9.

In this architecture, the Continuous Query Processing Component can query both stream and relational data. Streaming inputs come from the three different data sources: CCTV images, audio data and TRIPS [13] [14] data. The data at this stage is supposed to be processed and well structured. We expect to have data as Comma-Separated Values (CSV) or Extensible Markup Language (XML) files before performing queries. The Historical Database is another input in the form of the relational data. Historical data is important to provides the traffic prediction information. Streaming Result is the data stream containing permanent query answers while the Stored Result contains temporary query answers. The Scratch Store component stores the derived data from input for further processing. This architecture is our first idea to process continuous transportation queries and will be adjusted to adapt with further findings.

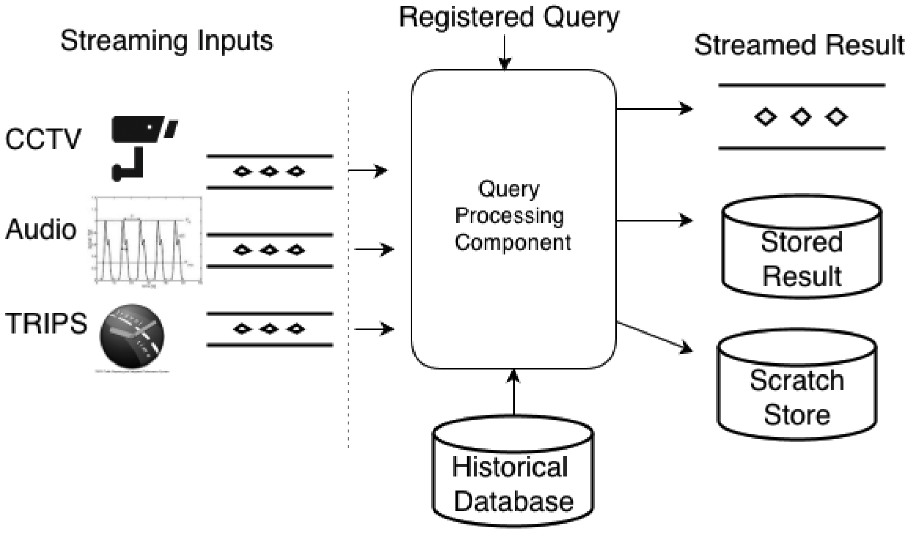


Figure 9: An Architecture for Processing Continuous Transportation Queries

6 Conclusion

In this paper, a case study of using acoustic data as a source for traffic monitoring in an ITS is undertaken. We found for free-flow conditions, the dominant frequency range is in the range of 800Hz-1.25Hz. Energy emphazization of congestion where vehicles move slowly is in the range of 200Hz-500Hz. Applying frequency filtering on relevant ranges, we can classify acoustic patterns of different traffic conditions from audio waveforms.

For the future works, we want to design and develop a prototype to process transportation queries. This prototype aims to retrieve precise and timely information from available transportation data sources. We have generated the sample audio data set in this study and proposed some ideas of applying streaming processing in continuous transportation queries.

Another potential research direction is to use mobile data for transportation purposes. Smartphones today are equipped with various powerful and low cost sensors that can provide accurate traffic information. A combination of different mobile data types (GPS data, video data and acoustic data) to support transportation management systems is also a promising research path.

7 Acknowledgement

This work was supported, in part, by Science Foundation Ireland grant 10/CE/I1855 to Lero - the Irish Software Engineering Research Centre (www.lero.ie).

References

[1] L. Gasparini, E. Bouillet, F. Calabrese, O. Verscheure, B. OBrien, and M. ODonnell, System and analytics for continuously assessing transport systems from sparse and noisy observations: Case study in Dublin, in 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2011, pp. 18271832

[2] B. Barbagli, L. Bencini, I. Magrini, G. Manes, S. Marta, and A. Manes, A Traffic Monitoring and Queue Detection System Based on an Acoustic Sensor Network, International Journal on Advances in Networks and Services, vol. 4, no. 12, pp. 2737, 2011.

[3] H. Doygun and D. Kuat Gurun, Analysing and Mapping Spatial and Temporal Dynamics of Urban Traffic Noise Pollution: a Case Study in Kahramanmara, Turkey, Environmental monitoring and assessment, vol. 142, no. 13, pp. 6572, Jul. 2008.

[4] A. Averbuch, V. A. Zheludev, N. Rabin, and A. Schclar, Wavelet-based Acoustic Detection of Moving Vehicles, Multidimensional Systems and Signal Processing, vol. 20, no. 1, pp. 5580, May 2008.

[5] A. Czyzewski and J. A. Ejsmont, Validation of Harmonoise/Imagine Traffic Noise Prediction Model by Long Term Noise and Traffic Monitoring, in Joint Baltic-Nordic Acoustics Meeting, 2008, pp. 17 19.

[6] N. Bhave and P. Rao, Vehicle Engine Sound Analysis Applied to Traffic Congestion Estimation, in CMMR/FRSM, 2011, pp. 59.

[7] G. J. P. John Patrick Kuhn, Binh C Bui, Acoustic Sensor System for Vehicle Detection and Multi-lane Highway Monitoring, U.S. Patent US57989831997.

[8] R. Sen, B. Raman, and P. Sharma, Horn-ok-please, in Proceedings of the 8th international conference on Mobile systems, applications, and services - MobiSys 10, 2010, p. 137.

[9] R. Sen, P. Siriah, and B. Raman, RoadSoundSense: Acoustic Sensing based Road Congestion Monitoring in Developing Regions, 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, pp. 125133, Jun. 2011

[10] Audacity Team, About Audacity[Online] Available: http://audacity.sourceforge.net

[11] World Bank website, Transport in East Asia and Pacific [Online] Available: http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/EASTASIAPACIFICEXT/EXTEAPREGTOPTRANSPORT/0,,contentMDK:20458737m̃enuPK:2069374p̃agePK:34004173p̃iPK:34003707t̃heSitePK:574066,00.html

[12] S. Babu and J. Widom, Continuous Queries Over Data Streams, ACM SIGMOD Record, vol. 30, no. 3, p. 109, Sep. 2001.

[13] Travel-time Reporting and Integrated Performance System [Online] Available: http://www.advantechdesign.com.au/trips

[14] P. R. Lowrie, SCATS, Sydney Co-ordinated adaptive traffic system: a traffic responsive method of controlling urban traffic, Roads and Traffic Authority NSW, Darlinghurst, NSW Australia, p. 28, 1990.

Traffic-condition Analysis using Publicly-Available Data Sets

Ulrich Dangel[1], Patrick McDonagh[2], Liam Murphy[1]

[1] LERO@UCD, School Of Computer Science and Informatics
University College Dublin, Belfield, Dublin 4, Ireland.
ulrich.dangel@ucdconnect.ie
liam.murphy@ucd.ie

[2] LERO@DCU, School Of Electronic Engineering
Dublin City University, Glasnevin, Dublin 9, Ireland.
patrick.mcdonagh@dcu.ie

Abstract

In this paper, we introduce some Dublin-specific public traffic data sets and analyse traffic data by linking it with other, non-traffic related datasets. We explain irregularities in observed journey times with weather phenomenas, public events and public holidays. We discuss how the timing of different weather phenomenas influences the observed journey time. By combining different data sources, we can provide reasoning for observed journey times which can be used to explain unexpected traffic patterns, improve capacity planning and aid with other traffic engineering tasks.

Keywords: ITS, Data Analysis, Ireland

1 Introduction & Related Work

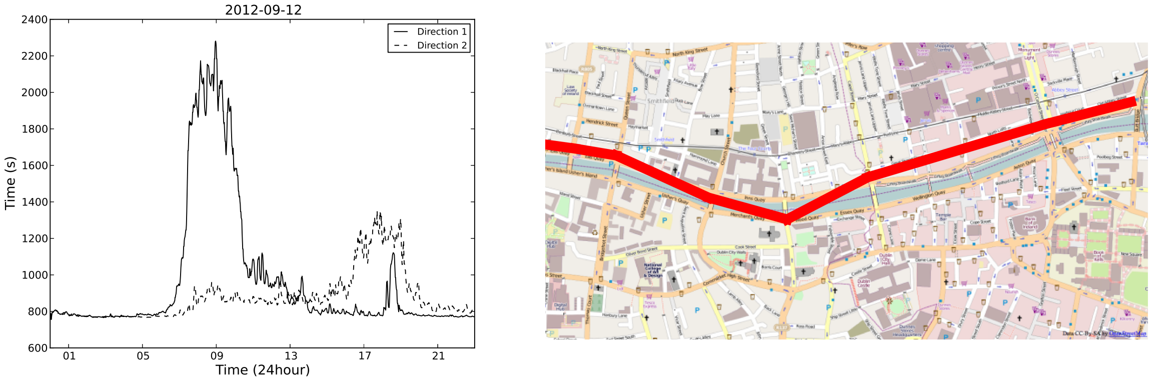
Dublin is ranked as the 16th most congested city in Europe which is worse than London and Munich [1]. As a method to control congestion Dublin City Council employs the Sydney Coordinated Traffic System (SCATS) [2] to handle traffic in the city. In 1989, Dublin installed SCATS [3] to monitor and control traffic; the overall monitored junctions increased during the years from 170 in 1997 [4] to more than 700 currently [5]. The overall increase in monitored junctions in Dublin emphasises the importance and necessity of an ITS.

An Intelligent Transportation System (ITS), such as SCATS, provides traffic engineers and system operators with the necessary metrics, such as congestion level [3], to control and monitor the traffic of a road network. The Irish National Transport Authority suggests the use of an ITS for controlling and monitoring large road networks to “enhance operational efficiency and driver information” [6].

The data provided by an ITS helps to monitor and detect problems such as traffic congestion or hotspots. As an ITS is a domain specific monitoring and control system, it does not aide in finding the cause of a problem as of result of events outside of the traffic domain, such as bad weather or large public events. By linking different data sources together, we can find what caused the problem, even if it was caused by events external to the domain.

The *intelligent* part of an ITS relies heavily on the data and data sources [7]. By using, for example, multiple traffic-related input sources such as induction loops, GPS probes [8] or ultrasonic sensors, a *multisource-driven ITS* can reduce the deployment costs by combining different sensor types or increase the monitored area by leveraging floating car data for cost-efficient traffic monitoring [7].

Understanding what caused a specific delay or congestion by combining and linking traffic and non-traffic related data sources was discussed in [9] with the focus on the semantics of the data sources. By



(a) Estimated Journey Time for Two Directions

(b) Route with wrong data point

Figure 1: Examples of data from Dublinked.

developing ontologies, they successfully linked different data sets such as roadworks and maintenance data with public transport data.

In our research, we focus on a general approach for monitoring a distributed system by using different data sets and combine these with other, not directly-related, data sources. Our main research goals are 1) finding irregularities in time series data for a distributed system and 2) helping diagnose irregularities by linking different data sets together. We are using traffic data as an example of an distributed system as such a cyber-physical imposes additional difficulties such as noisy data, regulation, privacy issues as well as communication issues.

Detecting and finding irregularities in time series data offers a broad range of applications such as 1) detecting defective sensors 2) data cleansing 3) verifying simulations and traffic models 4) supporting automated incident detection on traffic data by taking additional information into account.

2 Data Sets

There are a wide variety of public data sets available, from financial data to public health data. In this section, we will discuss some publicly available, local (Irish) data sets, which we will combine later to explain anomalies in traffic data.

Even though not all of the data sets discussed here are directly related to traffic, we will show how the combination and fusion of different data sources can explain traffic patterns and allow for a better understanding of the data and traffic conditions.

2.1 TRIPS Data from Dublin City Council

The Travel-time Reporting and Integrated Performance System (TRIPS) integrates with SCATS and is used to analyse and predict travel times between junctions monitored by SCATS. Dublin City Council publishes the data in real-time to Dublinked [5], a local website focused on publishing local, public sector data.

The published TRIPS data consists of the 1) locations of junctions monitored by SCATS, 2) a collection of defined routes and 3) journey time estimates. A route is a ordered list of links and a link is a ordered pair of positions associated with a junction. The journey time estimates are provided for each individual link of a route and for both directions. The explicit distinction between the travel directions can simplify the data analysis, e.g to compare morning and evening commuting, as shown in Figure 1a. This figure highlights the difference between the two directions, Direction 1 is primarily used in the morning, while Direction 2 is mostly used in the evening. Without this distinction, the spike at around 18:00 for Direction 1 would not be visible.

Route	Dublin City Council	Provider A	Provider B
9	12 minutes	11 minutes	13 minutes
11	5 minutes	6 minutes	5 minutes
24	14 minutes	15 minutes	13 minutes
30	14 minutes	13 minutes	12 minutes

Table 1: Estimated journey time comparison between TRIPS data and commercial routing services

2.1.1 Quality

The resolution of the TRIPS data is of a high granularity, as estimates are provided for the complete route as well as for individual links. The data is updated every minute, which allows for the detection of small, fine-grained patterns.

For a small subset of the routes, we verified the estimated travel time with the travel times from proprietary routing services, as shown in Table 1. The estimated times were similar during normal, non-congested times, i.e. in the early morning or evening. To compare and verify the provided estimates during congested time we would have to compare the estimates with real journey times captured by GPS probes [8] or by comparing it to other data, for example Dublin Bus data.

2.1.2 Limitations

Based on this data set, Dublin City Council currently (as of December 2012) monitors at least 1277 junctions with SCATS, though only 717 junctions have an explicit position, while others either have only street names or generic descriptions like *ROAD2*. This is only a minor issue as each link of the provided routes has a start and an endpoint.

Some of the 45 defined routes in Dublin City also have minor inconsistencies, such as alteration between the sides of the river Liffey, as shown in Figure 1b, or only consist of one link and other minor issues.

2.2 Journey Times from South Dublin County Council

South Dublin County Council (SDCC) publishes journey times on their website for 42 routes. The published journey times consist of a route name, average journey time, journey time and average speed. The *route name* is a pair of two canonical names; identifying the start and end point without an exact position. *Average journey time* is the average, historical, estimated time needed to traverse the route while *journey time* is the current estimated time. It is unclear however, how the journey time and average speed is calculated, e.g. by speed monitoring, induction loops or vehicle tracking with CCTV.

2.2.1 Quality

Depending on the time of day, the journey times are updated with a frequency ranging from five minutes up to multiple hours. Due to this delay, the data has a high variance between each of the data points. The published routes also don't contain enough information, as the exact route is not provided but only a start and end name, e.g. *Heuston Station to Palmerston*.

2.2.2 Limitations

SDCC does not publish the data in a machine readable format but on a website targeted for end-users [10], e.g. people travelling alongside one of the specified routes. In order to acquire the provided data, we need to scrape the website and extract the information. The calculations for a route may also be provided at different times, i.e. comparing routes to each others imposes additional challenges due to temporally unaligned data.

2.3 Weather data

The National Oceanic and Atmospheric Administration (NOAA) publishes weather information for all airports in the standardised Meteorological Aviation Routine Weather Report (METAR) format. METAR is a file format for reporting weather information which is widely used in aviation and meteorology.

METAR contains standard weather information like temperature, wind speed and weather phenomena as well as other aviation specific information like ICAO (International Civil Aviation Organization) code, visibility and cloud conditions. As METAR data is used for aviation purposes, the data can be considered to be of high quality.

We decided to use the weather data recorded from Dublin Airport as it is in close proximity to Dublin City Centre. While manually validating extraordinary weather phenomena (snow on January 22nd 2013) in Belfield, Dublin , we noticed a discrepancy in the recorded data even though the linear distance to weather station is approximately 13 kilometers. This can be mitigated by using more weather stations, providing increased local coverage.

2.4 Comparison

We choose to use the provided TRIPS times from Dublin City Council instead of SDCC as the TRIPS data has a higher resolution and is more complete, i.e. it specifies the actual route and not only an imprecise start and end name. The TRIPS data is also updated more regularly than the SDCC journey time data. Even though we are currently focusing on the TRIPS data, we may integrate the SDCC journey time data to get a better overview of the greater Dublin area by combining both data sets.

3 Data Processing

In this section, we discuss an subset of necessary steps to analyse different data sets, how to combine different data sets, as well as different ways to analyse the used TRIPS data.

3.1 Temporal Alignment

Different data sets can have different time resolutions, e.g. the METAR data is provided half-hourly, while the TRIPS data from Dublin City Council is provided minute-by-minute. Other data sets may have no fixed time resolution or specify a time range, e.g. a concert from 19:00 to 21:00.

In order to find corresponding sensor readings from other, temporally unaligned sensors, we need to map the data from one time specification to another. This mapping depends on the sensor type. Observations for different phenomena have a different underlying model; traffic behaves different than temperature, for example. Based on the underlying model, different methods may be employed.

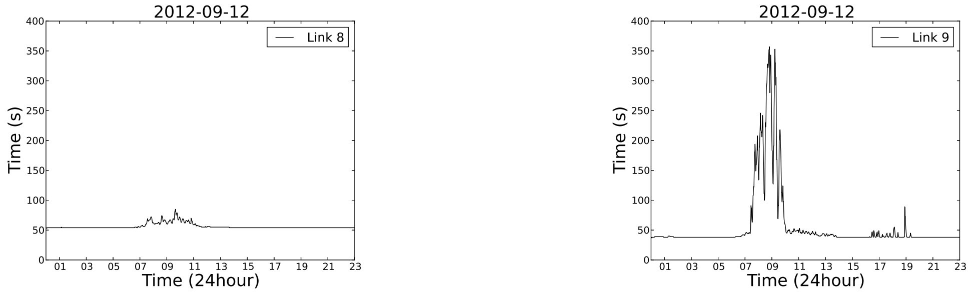
3.1.1 Interpolation

Depending on the nature of underlying data, a simple linear interpolation between the measured data points or more advanced interpolation based on a model may be feasible. For air temperature, we can use a model-based interpolation to reduce the average error rate and improve the overall findings [11].

3.1.2 Time range

While air temperature can easily be interpolated, other data such as weather phenomena is difficult to interpolate. Instead of interpolating the data points themselves, we can transform the measurement to a time range by combining it with the previous and next data point. Equation 1 and 2 shows a simple method to define a new range based on the predecessor and successor of a data point, where t_s is the new start point, t_e the new end point and t_n the time of the nth measurement

$$t_s = t_n - \frac{t_n - t_{n-1}}{2} \tag{1}$$



(a) Estimated Journey Time for Link 8.

(b) Estimated Journey Time for Link 9.

Figure 2: Comparison of Journey Time estimates for different links.

$$t_e = t_n + \frac{t_{n+1} - t_n}{2} \tag{2}$$

Depending on the observed weather phenomena, the formulas can be adjusted to take additional information or transformations into account and dynamically adjust the period on the next and previous conditions.

3.2 Micro versus Macro Analysis

The TRIPS data from Dublin City Council provides journey time estimates for individual links, as well as for complete routes. This allows us to compare congestion propagation as well as how merging multiple routes onto the same link influences traffic.

Figure 2 shows two links on the same route. Figure 2a (Link 8) shows a link with only one route passing through, while Figure 2b (Link 9) shows a link where multiple routes merge. The journey time on Link 9 may be higher than on Link 8 as after Link 9 three routes are merged traveling towards Dublin City. This means that traffic signals are adjusted to merge these three major routes together and creates congestion, providing a possible reason for the increased journey time.

Even though micro analysis can provide a good insight into specific routes or behaviour, we are currently focused on macro analysis to provide a good overview of pattern throughout a city. With the combination of macro and micro analysis, we can provide a good explanation on specific time pattern caused by hotspots.

4 Patterns

In this section, we will discuss some of the patterns (observed journey times) we detected while verifying and analysing the TRIPS data from Dublin City Council. Currently, this analysis is performed manually, future work will extend this analysis by automating this approach using stochastic methods. We will provide an explanation for these patterns by linking it with external, non-traffic data sources such as weather data and public event data.

4.1 Bank Holiday

Figure 3 shows the journey time for three consecutive Mondays for a specific route. The last Monday in Figure 3 has a significantly lower average estimated journey time than the previous Mondays, as October the 29th 2012 was a bank holiday in Ireland.

The reported journey time for the public holiday matches expectations as the observed route is a commuter route into Dublin city. The reduced commuter traffic, due to closed local offices and businesses, reduce the overall journey time. The slightly increased journey time in the morning could be ascribed to traffic travelling into the city center for shopping and/or to the restrictions imposed by the Dublin Marathon, occurring at the same time.

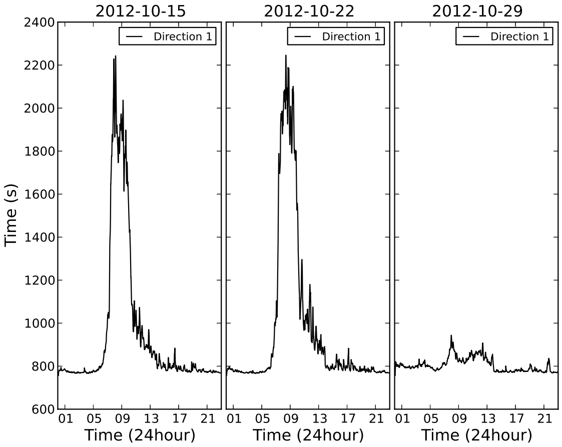


Figure 3: Comparison of journey times between a bank holiday and regular days.

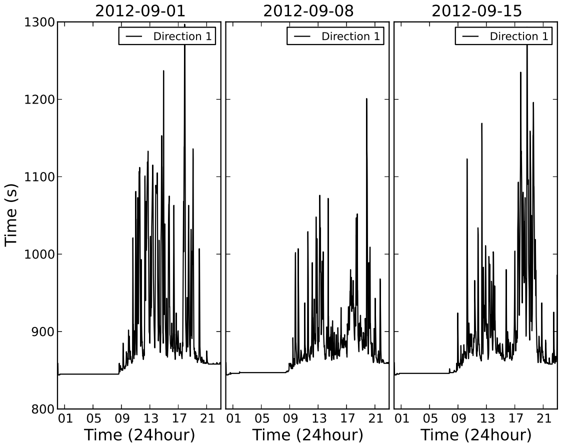


Figure 4: Different journey times for public events located near a route.

4.2 Public Event

Bank holidays produce a simple traffic pattern, as the majority of the workforce are not required to work. A slightly less obvious pattern, caused by large public events, is shown in Figure 4. The estimated journey times are for a route in close proximity to the Aviva Stadium, a local sports stadium with a capacity for over 50,000 spectators.

The traffic pattern shown in the first chart is caused by an sold out American Football match in the Aviva Stadium. To cope with the traffic, the police closed some roads at 09:30; at 12:00 the local police closed more roads to cope with the spectators [12]. The temporary reduction of the estimated traffic time could be explained by the partial road restrictions imposed by local police. Directly after the event, at approximately 17:30, there is a large increase in journey time as spectators leave the premises.

The second graph in Figure 4 shows an typical Saturday with no events in the Aviva Stadium. The relatively small and short increase around 13:00 could be ascribed to smaller events in close proximity to the route such as TedX (in the Bord Gais Energy Theatre), or shoppers travelling into the city centre.

The last chart in Figure 4, shows the estimated journey times caused by an sold-out concert at the Aviva Stadium. The gates opened at 17:00 which is visible in the journey time estimates. The decreased journey time after 13:00 can be ascribed to local police authority closing roads around the sport stadium [13]. The morning traffic pattern resembles roughly the previous, normal week but the evening traffic differs a lot due to the sold-out event in the evening.

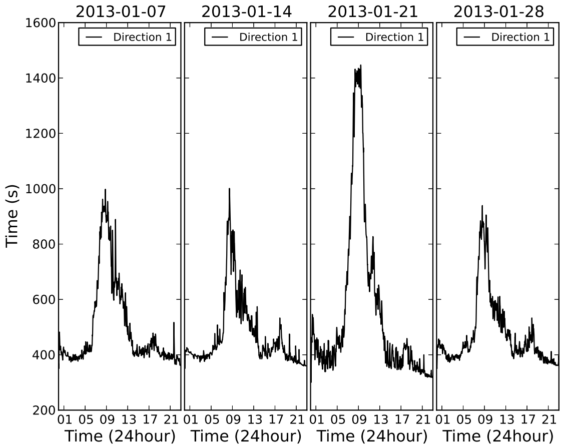


Figure 5: Influence of weather on journey times.

4.3 Weather

Another pattern we detected is that caused by severe weather. We chose a route from Dublin North to Dublin City Centre as representative of typical commuter traffic. Figure 5 shows journey times on consecutive Mondays; even though on all the Mondays, some weather phenomena was recorded, different journey times can be observed.

For January the 7th (first chart), 14th (second chart) and 28th (fourth chart) *rain* was registered by the weather station in Dublin Airport. On January 7th, rain was registered throughout the whole day while, for January 14th rain was only registered up until 6 in the morning and later in the afternoon. The increased travel time persists in the first chart over a longer time period than in the second chart. Compared to these two, quite similar charts, rain on January the 28th was first registered at 9 in the morning. The increase, compared to previous Mondays, happened later, around 9 when rain was detected.

On January the 21st *snow* was registered throughout the whole day. The severe and high increase in journey time can be attributed to snow as it is an uncommon event in Dublin. The increased journey time was not only higher than to the other days but it also lasted for a longer period.

4.4 Findings

We provided causes for specific traffic behaviour and showed that, by using non-traffic related data sources, we can aide and provide reasons for specific irregularities in journey times.

The observed journey times had different characteristics; (1) on the public holiday there was a strong decrease in the estimated journey time throughout the whole day due to reduced overall traffic levels, (2) how large public events influenced the traffic for routes close to the premises but how mitigation employed by the local police helped to reduce the observed travel time and (3) how different weather phenomena, as well as their timing and duration, have a large influence on the journey time.

5 Conclusion and Future Work

We introduced and assessed different, publicly available, local data sets for traffic information and other non-traffic related data sets. We verified the traffic data by explaining and cross-referencing observed journey time patterns with other data sets. We have shown the influence of bank holidays, large public events and weather phenomena such as rain and snow on journey time.

Our current, manual, approach for detecting abnormal behaviour is driven by visualising journey time estimates. This manual process has been proven to be valuable to verify and understand unknown data sets. In order to scale and process a large amount of data, we plan to automate the process of finding irregularities in time series data by creating a model for the data.

In order to find an appropriate model, we plan to use different statistical analysis methods to detect, for example, periodic, shifted data. The analysis of time series data can also help to classify unknown behaviour. We plan to use traffic related data sets as an use case for our approach as the highly distributed nature introduces additional challenges for the data quality and the amount of incurred data.

Our aim is to find a general approach for analysing time series data and cross referencing it with data sets from other domains. As distributed systems grow larger and become more complex, there is a need for automatic detection of irregularities and to fuse performance data with other data sets. We are currently using traffic data as a city resembles a distributed system in the physical world. By cross-referencing data sets, we showed that our approach can provide a better understanding of the observed system and support resource and capacity planning, as well as root cause analysis.

Acknowledgments

This work was supported, in part, by Science Foundation Ireland grant 10/CE/I1855 to Lero - the Irish Software Engineering Research Centre (www.lero.ie)

References

[1] TomTom. (2012) TomTom European Congestion Index. [Online]. Available: http://www.tomtom.com/lib/doc/congestionindex/2012-1003-TomTom-Congestion-Index-2012-Q2-europe-mi.pdf

[2] A. Sims and K. Dobinson, "The sydney coordinated adaptive traffic (scat) system philosophy and benefits," *Vehicular Technology, IEEE Transactions on*, vol. 29, no. 2, pp. 130–137, 1980.

[3] M. Dineen, "Real-Time Display of Dublin Traffic Information on the Web," Master's thesis, Trinity College Dublin, 2000.

[4] M. Fellendorf, "Public Transport Priority within SCATS - a simulation case study in Dublin," in *Institute of Transportation Engineers 67th annual Meeting*, 1997.

[5] Dublin City Council. (2013) Journey times across Dublin City. [Online]. Available: http://dublinked.ie/datastore/datasets/dataset-215.php

[6] National Transport Authority. (2011) Greater dublin area draft transport strategy 2011-2030. 2030 vision. [Online]. Available: http://www.2030vision.ie/downloads/files/en/final/draft_strategy.pdf

[7] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 4, pp. 1624 –1639, dec. 2011.

[8] J. Aslam, S. Lim, X. Pan, and D. Rus, "City-scale traffic estimation from a roving sensor network," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. ACM, 2012, pp. 141–154.

[9] F. Lécué, A. Schumann, and M. Sbodio, "Applying semantic web technologies for diagnosing road traffic congestions," *The Semantic Web–ISWC 2012*, pp. 114–130, 2012.

[10] South Dublin County Council. (2013) Traffic and Travel Information. [Online]. Available: http://traffic.southdublin.ie/car_journ/journey.aspx

[11] C. Willmott and S. Robeson, "Climatologically aided interpolation (cai) of terrestrial air temperature," *International Journal of Climatology*, vol. 15, no. 2, pp. 221–229, 1995.

[12] Ireland's National Police Information. (2012) Garda Information for Notre Dame -V- Navy. [Online]. Available: http://www.garda.ie/Controller.aspx?Page=9798

[13] ——. (2012) Garda Information for Lady Gaga Concert. [Online]. Available: http://www.garda.ie/Controller.aspx?Page=9879

Making a Case for an Irish500 List

Brett A. Becker [1], John Regan [2], Michael Salter-Townshend [3], Kevin Casey [4]

[1] College of Computer Training, Dublin 2, Ireland
brett.becker@cct.ie

[2] Helsinki Institute of Information Technology, Aalto University, Helsinki, Finland
john.regan@hiit.fi

[3] Complex and Adaptive Systems Laboratory, University College Dublin, Dublin 4, Ireland
michael.salter-townshend@ucd.ie

[4] School of Computing, Dublin City University, Dublin 9, Ireland
kcasey@computing.dcu.ie

Abstract

Ireland is uniquely positioned in the global High Performance Computing arena. Among the reasons for this are a small population coupled with diverse groups of world-class academic institutions, well-funded government bodies, and world-leading research groups. In addition, Ireland is a location of choice for the European/EMEA/Global headquarters of a large number of multinational computing and technology companies. Nonetheless, the Irish High Performance Computing (HPC) landscape is difficult to define domestically and even more difficult to place amongst the international community. This is largely due to the lack of a central repository of Irish HPC resources, cataloguing their capabilities, and application areas. A natural way to organise such resources is to rank them in terms of performance. This format also serves to command wider interest from outside the HPC community. As HPC resources have become a strategic asset, the ability to quickly identify them qualitatively and quantitatively is a powerful resource, particularly when this information can be searched, partitioned and tailored for specific uses. Projects such as the Top500 list have served this purpose for years, but only for the countries with the budgets and resources to place on such lists with a statistically significant number of systems.

This paper makes a case for an "Irish500" list. The distinguishing feature of this list is that all systems must be geographically located in Ireland. The two primary questions this project seeks to answer are: What is the landscape of HPC in Ireland today? Where does this landscape fit globally? The mission of the Irish500 list is the promotion, advocacy and advancement of HPC in Ireland. We persist with the convention of using 500 in the name for consistency, as have most other lists since the Top500 started; we are not aiming to create a list that actually comprises as many as 500 machines. The list ranks Irish HPC installations using the Linpack performance benchmark which is also used by the Top500 and other lists. In this paper we also make the case for the introduction of a flops/watt metric, used by the Green500 list, to rank systems in terms of computational energy efficiency. We discuss the motivation behind and benefits of the Irish500 list and explain its relation to existing lists. We then explore work based on, and interest in, other HPC lists. Finally, we describe the structure of the proposed list, including benchmarks and metrics used, along with a schedule of releases and anticipated participant sectors. We then conclude with a discussion on the implementation of the list and the irish500.org website.

Keywords: Top500, Green500, Irish500, Supercomputing, High Performance Computing

1 Introduction and Motivation

The Top500 Supercomputers (www.top500.org) has been maintaining a list of the most powerful computers in use for over 20 years. The Top500 has proven to be extremely important for several reasons. At a very high level, HPC resources have become a strategic asset. Being able to quickly identify them qualitatively and quantitatively is a powerful resource, particularly when this information can be searched, partitioned and customised for particular uses. At a very basic level, it has been a driver for HPC procurements and provides funding attractiveness and enhanced recruitment power. These benefits apply not only to those at the top of the list but arguably for most sites that rank on the Top500. It has been shown that funding attached to Top500 ranking systems has a direct link to their performance, and conversely that high ranking attracts more funding and is also associated with an increase in the number of publications [1]. The list also has value in a public relations role, with journalists in many major newspapers reporting the latest list releases, drawing public attention and opinion to HPC in general. Again there is a link to funding in this regard, as politicians and policy-makers have taken an interest in the Top500 [2]. In practice, the Top500 has influenced the HPC industry to the extent that new supercomputer installation announcements are often clustered around the releases of list updates in June and November of each year [3]. The Top500 has even trickled down into general public awareness to the point that a recent *New York Times* article featuring a Top500 founder discussed building a Top500 machine using networked iPads [4]. HPC capabilities have even become a source of national pride [5].

In 2006 the second HPC ranking list, the Green500 (www.green500.org) was announced, ranking installations using the flops/watt metric [6]. The Green500 was a direct response to both the public interest in environmental friendliness, and more directly, to the rapidly growing power consumption problem in HPC. A third global list, the Graph500 was announced in 2009 (www.graph500.org). This was a response to criticism that the Linpack [7] benchmark which is utilised by the Top500 is not the most representative benchmark for predicting the performance of the actual applications that HPC installations execute. We discuss this further in Section 4.2. The Top, Green and Graph lists share one thing in common – their scope is limited to the top 500 performing systems *globally*. The only list known that ranks a different domain is Top Supercomputers India which ranks only those systems located in India [8].

1.1 Motivation for an Irish Supercomputer List

Ireland is uniquely positioned in the global HPC arena. Despite a small population (less than seven million, including Northern Ireland) there are large, diverse groups of world-class academic institutions, well-funded government bodies, and world-leading research groups. Moreover, Ireland is a favoured European/EMEA/Global headquarters location for several multinational computing and technology companies. Ireland has hosted eleven Top500 ranked installations, with ten in the past decade. Figure 1 charts these systems and their rank history. Table 1 shows installation owner/site, brief architecture details, and Top500 information for these systems.

This history is evidence of Ireland’s ability to build and attract world-class HPC facilities. However, this hardly approaches a complete picture of HPC in Ireland, particularly at the present time. Many world-class systems in Ireland are not represented here simply because they fall short of the Top500 global ranking. In addition, five Irish Top500 systems have been decommissioned. It is also known that there are Irish companies and institutions that currently possess hardware that will rank highly domestically, but their performance data is not publicly available at present, or benchmarking of any kind has not yet been done. It is a motivation of the Irish500 list to compile a more representative and current picture of the Irish HPC landscape.

As the Irish HPC infrastructure is difficult to define domestically, it is even more difficult to place it amongst the international community. This is largely due to the lack of a central repository of Irish HPC resources and their capabilities. A natural way to organise such resources is to rank them in terms of performance or capability. Such repositories (such as the Top500) have served many purposes for years, but primarily for the countries with the budgets and resources required to place systems on them. The Top500 can yield useful, detailed and important information for countries such as Canada for example, which places seventh globally in terms of HPC resources. Such information is used for many purposes, including national strategy and budgetary decisions, funding applications, and even matters such as a sense of community amongst computer scientists. Currently Ireland does not have enough representation on the Top500 to serve these purposes, nor has it a domestic platform which can do so. It is a chief motivation of the Irish500 to provide such a platform for Ireland. First and foremost, the Irish500 seeks to answer:

- *What is the landscape of HPC in Ireland today?*
- *Where does this landscape fit globally?*

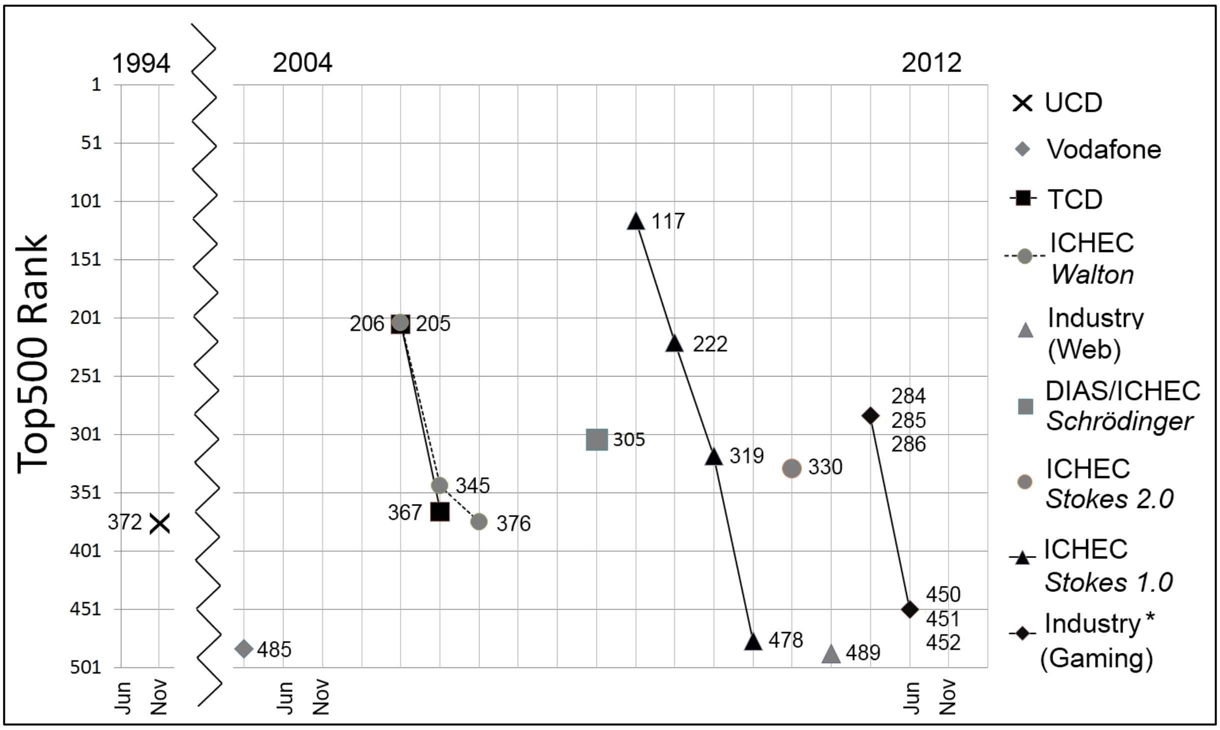


Figure 1: Ireland's history of Top500-class HPC installations. *(three identical clusters)

Installation	Max Top500 Rank	Top500 Dates	Hardware Specification
ICHEC	117	11/08 – 6/10	SGI Altix ICE 8200EX, Xeon 4C 2.8G
ICHEC	205	6/05 – 6/07	IBM eServer Opteron 2.4G, 1G Eth
TCD	206	6/05 – 11/05	IBM eServer Opteron 2.4G, Infiniband
Industry (Gaming)*†	284/285/286	11/11 – 6/12	HP Cluster Platform 3000 BL460c G7, Xeon X5660 6C 2.80G, 10G Eth
DIAS/ICHEC	305	6/08	IBM BlueGene/P
ICHEC	330	11/10	SGI Altix ICE 8200EX, Xeon 6C 2.66G
UCD	372	11/94	MasPar MP-2216
Vodafone	485	6/03	HP SuperDome 875 MHz, HyperPlex
Industry (Web)†	489	6/11	HP DL160 G6, Xeon X5650 2.66G, 1G Eth

Table 1: Site and architecture information for Ireland's Top500-class HPC installations in increasing order of highest Top500 rank achieved.
***(three identical clusters) †Anonymous submissions to Top500 list**

The only country known to have a dedicated domestic list is India [8]. Despite having only eight systems on the latest Top500 list, the December 2012 Top Supercomputers India list has 26 operational systems, giving a robust picture of the Indian HPC landscape. Given the relatively good finances of the Irish research sector, and abundance of multinational IT companies, the Irish list could easily gain as many systems. Additionally, a relatively small population is seen as an advantage – in particular it brings easily identifiable and contactable HPC providers/users in addition to making comprehensiveness a realistic possibility. If each university, institute of technology, and a handful of multinationals were to enter just one system each, the number of installations on the Irish500 could exceed 30. If submissions are made by individual faculties, schools, research groups and smaller companies (not to mention multiple submissions by each) the number of systems becomes substantial enough to be valuable for substantial analysis, including market trends and hardware/interconnect usage, as well as application information. We explore this further in Section 4.3 and present a feasibility study demonstrating that the Irish500 list could realistically equal or surpass the Indian list, both in system numbers and performance in the short-term. This would be a positive result for the Irish HPC community, and a major step towards addressing the questions posed above.

Another motivating factor in maintaining an Irish list is to promote accuracy and representation. In November 2011, Amazon Web Services benchmarked a system in the U.S. made of 1062 EC2 cc2.8xl (Cluster Compute Eight Extra Large) instances. The result was 240 Tflops, placing it at #42 on the June 2011 Top500 list, and currently at #102 [9]. This cluster contained 17,024 cores with 66 TB of RAM. In 2012 Amazon announced that these same instances were available in their Irish (EU-West) region [10]. The combined capability of these instances in the Irish datacentre is not currently known, but unless the number of machines is significantly less than that in the U.S., the Irish Amazon region could also place in the Top500 list and very possibly top the Irish500.

Additionally, the Irish Amazon region has been used in two global projects conducted by the software firm Cycle Computing. The first, *Nekomata*, is a 30,000 core homogeneous effort consisting of machines in three Amazon regions. The exact Irish contribution is known, therefore allowing a very reliable estimate of the performance of the Irish machines involved [11]. The second, *Naga* is a 50,000 core, seven region effort. Unfortunately the system is made of highly heterogeneous instances and although the number of Irish machines is known, the breakdown of instance types is unknown and therefore a reasonable performance share estimate is not feasible [12].

It is widely believed, based on experiments that probe EC2 and calculate numbers of racks by various means, that the Amazon EU-West (Ireland) region is the second largest of the seven EC2 regions, second only to US-East (Virginia) [13]. The fact that these resources exist in Ireland, and have paying customers conducting hard science [11] [12] [14] is encouraging, yet the true capabilities of these resources remain unknown. It is a motivation of the Irish500 list to identify such capabilities, while respecting any installation owner’s right to anonymity and/or privacy of certain specifications when required.

Finally, the Irish500 list will help to fill the void left with the closure of Grid Ireland at the end of 2012 [15]. Amongst many hardware based services, Grid Ireland also functioned as a central networking point for the Irish HPC community, something that the authors believe needs to be fostered to maintain that sense of community and forward national momentum. It is believed that the Irish500 can help serve as one such central networking body.

2. Existing Lists and Projects

There are four related lists now in operation: The Top500, Green500, Graph500 and Top Supercomputers India. In addition, the HPC Challenge Benchmark serves a similar purpose to the traditional list format, as we discuss below. It is included here, as together these are the only known implementations of HPC rankings.

2.1 Top500

The Top500 list ranks the top 500 supercomputers currently in operation. It was established in 1993 and is updated twice per year – in June coinciding with the International Supercomputing Conference (always held in Europe) and in November coinciding with the ACM/IEEE Supercomputing Conference (always held in the US). The Top500 continues to grow in prominence, attracting widespread media and public attention, not to mention serious attention from researchers, universities, politicians and funding bodies.

The Top500 uses a single benchmark, Linpack, most often in the form of HPL, a portable implementation of the high-performance Linpack benchmark for distributed-memory computers [7]. This benchmark calculates the floating point operations per second (flops) achieved during the solution of a dense system of linear equations. The actual metric is Rmax in Tflops, where Rmax is the maximum sustained performance during the benchmark run. This is related to Rpeak, the theoretical maximum performance, though the relationship varies from installation to installation. Summing all 500 Rmax and Rpeak values in the latest Top500 list reveals an Rpeak/Rmax ratio ≈ $\sqrt{2}/2$. This is useful to estimate the Rmax value of a machine from the Rpeak, in cases where a machine has not yet done any benchmarking but has a known theoretical performance. Admittedly Linpack has drawn some criticism of late, for not being conceptually representative of real-world applications executed by HPC systems [16]. Nonetheless, it is simple, relatively easy and cheap to run, is very scalable, and has no direct replacement as of yet [17]. Most importantly, the Top500 has grown in influence year-on-year since inception, and continues to use Linpack, with no sign of this changing in the near future [2].

2.2 Graph500

The Graph500 list was started in 2010 as a response to the criticism of Linpack [16]. It utilises a suite of data-intensive benchmarks, using a metric of TEPS (Traversed Edges Per Second) which is designed to emphasise the importance of the communication network as well as the number-crunching capabilities of an installation. As of November 2012, there are 124 systems on the Graph500.

2.3 Green500

As the power consumption of supercomputers has risen well into the megawatts, energy efficiency has come to the forefront of HPC. The Green500 list ranks the current Top500 installations not in terms of flops but flops/watt [6]. The list is generated with Top500 data and always has 500 systems listed – a list equivalent to the Top500 except utilizing a different ranking metric.

2.4 Top Supercomputers India

Top Supercomputers India (TSI) was started in November 2008, but now lags the release schedule of the Top, Graph and Green lists by one month. This is most likely due to the possible overlap of Top500 and TSI installations (which is of course allowed, but needs to be consistent). It is the closest analogue to the proposed Irish500 list in that it is limited to installations in a specific geographical area, and performance need not be amongst the top 500 globally. The benchmark used, as for the Top500, is Linpack. The November 2012 TSI list has 27 ranked installations. One of the likely differences between the proposed Irish500 list and the TSI list is that all TSI installations, barring one from commercial giant Tata are owned or administered by national/governmental bodies or research/education/university institutions. Such a lack of industry presence would not be expected to characterise the Irish list, as almost half of Irish Top500 systems have been in the industry sector, including the fastest Irish installation ever. Section 4.3 shows that the proposed Irish500 list and the TSI are similar in number of systems and performance.

2.5 The HPC Challenge Benchmark

The High Performance Computing Challenge (HPCC) is a departure from the traditional “500-style” list [18]. It is a series of awards aligned to a suite of seven benchmarks with a similar aim to those of the Graph500 – to test a system’s performance with benchmarks that are representative of real-world HPC applications. It is worthwhile to note that Linpack (HPL) is one of the seven benchmarks in the HPCC suite, and that some of the others are closely related to it. Due to the nature of the benchmarks and the small number of awards, the HPCC looks poised to always be in the domain of the supercomputing super-elite, perhaps the global top 50 or so.

3 Related and Resulting Work

The existing lists have been the topic of, and source of data for, research and debate in the literature. Being updated twice per year, a substantial amount of data is compiled including where the installation is located, the installation performance history, architecture, manufacturer, processor family, age, benchmark performance, core/node/processor count, interconnect, power consumption, OS, accelerator information and possible application areas. This makes for a very good dataset upon which to identify current trends and predict future ones. Some of the latest data analysis from the Top500 was featured in the June 2012 cover story of Scientific Computing [19]. One of these insights revealed that the research and enterprise system technology gap is growing. Without such hard data, this would be impossible to demonstrate quantitatively.

In a 2009 study, the question “How high can a cloud computing service get in the Top500 list?” was addressed [17]. The conclusion was that cloud computing platforms (specifically Amazon EC2) were not yet mature enough for HPC applications. Nonetheless, this paper did give an indication that, at some point in the near future, cloud platforms would be viable for HPC applications. The authors' assertions were validated just two years later when Amazon had a cloud platform place at #42 on the Top500 [9]. In addition the paper demonstrated the viability of the economics involved in running the Linpack benchmark on pay-on-demand cloud platforms, with single runs of Linpack costing only tens of dollars. In this case the Top500 list was the perfect venue to see where a platform new to HPC/Supercomputing fitted, and to make predictions about the future of the platform in HPC. Similar stories are unfolding in the GPU/Accelerator arena, as the reality of systems running more than one million cores has arrived [20].

Other studies have sought to extract more qualitative data from the current lists including identifying and predicting trends in HPC. In [5] the authors identify invariant trends in Top500 data, predicting growth rates, and identifying limiting factors. On the Green500 front, recent forecasts indicate that the “performance at any cost” paradigm is no longer sustainable as HPC moves towards the exascale [21]. Such studies are valuable to the HPC community and in some cases have already served as indicators of things to come.

Top Supercomputers India has also spawned some work of its own, in the form of a project that collects additional information about the machines on the Indian list including numbers of jobs per month, how many cores jobs use, application categories, fault-tolerance, and other performance information. Currently, the preliminary outputs of this project are pending [8].

4 The Irish500 List

An advisory committee has been formed for the Irish500 list, including personnel at Irish and foreign bodies (for international and unbiased perspective). As efforts gain pace, the size and diversity of this group will be encouraged to grow. Having laid out the motivation and justification for an Irish500 list, we will now present initial work on the Irish500 project.

4.1 Mission Statements

The founding mission statements of the Irish500 follow, and will be regularly reviewed:

1. To form a central point for HPC installation operators and users in Ireland across all sectors, public and private, including (but not limited to) academic, research, industry and government.
2. To improve the awareness of HPC amongst the Irish public.
3. To represent and enhance the profile of the HPC landscape domestically and internationally.
4. To serve as a globally-facing resource representing Irish HPC, in keeping with both global and national conventions.

5. To identify current and emerging trends in Irish HPC, which can then be compared to those identified in global lists.
6. To provide a list whose statistics are representative of the Irish HPC landscape, for use by anyone for any purpose.
7. To maintain a list whose statistics are not dominated and skewed by a small number of systems at the top end – systems that also often feature expensive, exotic or custom hardware not available to most users of HPC systems.
8. To allow any university department, vendor, research group or company to rank on a list of peers, to quantitatively determine their rank, and qualitatively identify and analyse the properties and trends of similar systems, by similar organisations, with similar goals and means.
9. To provide a lifetime-long, not time-on-list-long history of system performance. This will be achieved by tracking system performance until decommissioning, not only until failure to attain a specified rank or performance mark.
10. To identify hardware and cooling trends that lead to cost effective performance and maximum energy-efficiency.

4.2 Choice of Metrics

The Irish500 will ultimately feature two separate rankings. The first and higher priority is based on computational performance. As the Irish HPC community becomes more involved, the second, ranking energy efficiency will be introduced. The energy efficiency metric will be flops/watt, measuring the average power consumption of a system while running at full capacity, identical to the Green500. The performance metric used by the Irish500 list will be that of the Top500 list which calculates system floating point operations per second (flops), using Linpack. The principal reasons for this are that it is cheap and simple to run, produces a single figure, is very scalable and has longstanding historical precedence. On a more technical level, we also choose Linpack because it gives a good indication of performance, despite some recent criticism. Most of this criticism has been directed at Linpack not stressing enough system components. Indeed, it was in response to this criticism that Graph500 was created. Although it is not the purpose of this paper to support or criticise either Linpack or proposed alternatives, a brief case to justify the use of Linpack as the Irish500 computational performance metric beyond the recent criticism is presented. This is done by demonstrating that good Linpack performance can be shown to correlate with good Graph500 performance (and vice-versa), within a very acceptable margin.

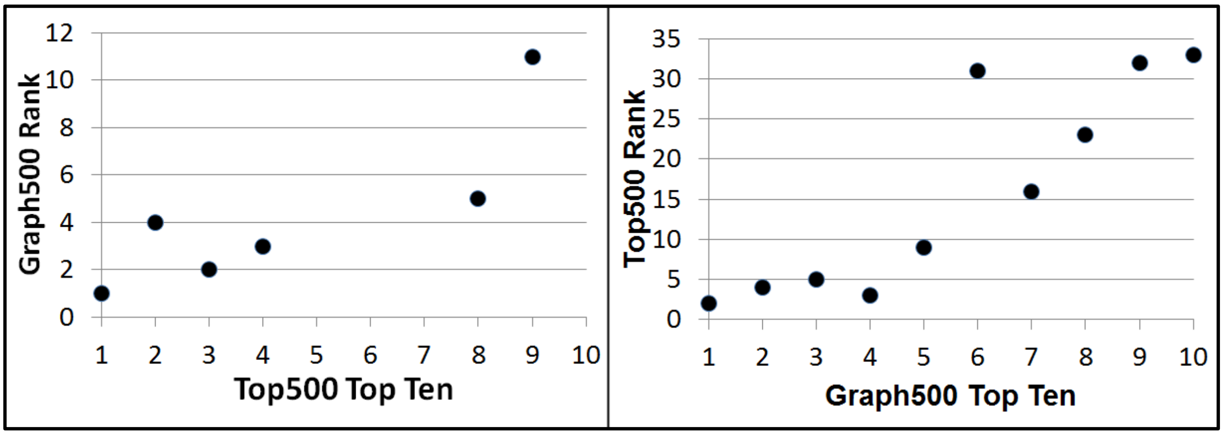


Figure 2: The top ten machines on the Top500 and Graph500 lists and their corresponding ranks on the other list (November 2012).

Figure 2 shows the top ten installations on the November 2012 Top500 and Graph500 lists, and their corresponding rank on the other list. All six machines in the Top500 top ten which entered the Graph500 benchmarks (not all did) are in the top eleven on the Graph500. Similarly, the top five

Graph500 machines are in the top ten on the Top500, and all top ten Graph500 are in the top 33 of the Top500. It is important to point out that this is not necessarily a like-for-like comparison, as machines may have run each benchmark under different configurations. Either way, it is interesting and important to see that good Top500 performance can be an indicator of good Graph500 performance, and vice-versa. This supports the decision making the cheaper, faster, simpler, and historically proven Linpack the benchmark for the Irish500. Finally, the Top500 website states “It is very unlikely that another benchmark will replace Linpack as basis for the Top500 lists in the near future” [2].

4.3 Moving Forward - Feasibility

A pre-release feasibility list as of February 2012 can begin to take shape now, based on information in the current and past Top500 lists and other publicly available information [22] [23] [11] [24]. It should be noted that a goal of the Irish500 is to develop direct contact with all installations that feature on future lists, therefore not relying on public information, but verifiable, direct-from-source information. The feasibility list is shown in Table 2, and was compiled with the sole purpose of determining the feasibility of the Irish500 project, having stated the motivation and justifications behind it.

Rank	Top500 Rank	Rmax (Tflops)	Installation	Hardware Specification
1 (**1**)	285*	64.9	Industry (Gaming)	HP Cluster Platform 3000 BL460c G7, Xeon X5660 6C 2.80 GHz, 10G Eth
1 (**2**)	286*	64.9	Industry (Gaming)	HP Cluster Platform 3000 BL460c G7, Xeon X5660 6C 2.80 GHz, 10G Eth
1 (**3**)	287*	64.9	Industry (Gaming)	HP Cluster Platform 3000 BL460c G7, Xeon X5660 6C 2.80 GHz, 10G Eth
2 (**4**)	489*	40.5	Industry (Web)	HP DL160 G6, Xeon X5650 2.66 GHz, 1G Eth
3 (**5**)	330* 117**	36.6	ICHEC – Stokes	SGI Altix ICE 8200EX, Xeon X5650 2.67GHz, ConnectX Infiniband DDR
4 (**6**)		31.9†	Cycle Computing – Nekomata	500 Amazon EC2 c1.8xl instances, 10G Eth
5 (**7**)		8.9	TCD – Lonsdale	Opteron 2.3GHz, 1232 Core, Infiniband DDR
6 (**8**)		8.9†	TCD – Kelvin	Intel 2.66Ghz, 1200 core, Qlogic Infiniband QDR
7 (**9**)		8.3†	TCD – Parsons	Intel 2.5Ghz, 1104 core, Voltaire Infiniband (QDR switching and Connectx DDR hosts)
8 (**10**)		5.1††	ICHEC – Stoney	Bull Novascale R422-E2, Intel Xeon X5560 2.8GHz, ConnectX Infiniband (DDR)
9 (**11**)		1.0†	TCD – Crusher	Intel2.5Ghz, 144 cores, Infiniband
Total		**280.7**		

Table 2: The Irish500 pre-release (feasibility) list as it stands in February 2013, using only publicly available information, all systems currently in commission. *Highest historical rank **Highest historical rank (Pre-upgrade) †Estimated Linpack Rmax ††CPU only (no GPUs)

With 11 systems overall (three identical), the performance ranges from 1 to 64.9 Tflops, and includes six past Top500 machines. The sectors span academia, industry, research, and government. With such a spread of performance and sectors already identified, a combination of a call for participation, public awareness, and the advisory committee contacting identified bodies, this list will surely grow. Only one university (TCD) features on this list, because they are the only one with publicly available figures. There is little doubt that UCD, UCC, UCG, NUIM, DCU, QUB and other academic institutions could significantly contribute to this list. In fact, it is known that most of these institutions have hardware in place that *will* rank highly in this list. Industry has a strong presence in the feasibility list with five installations. Certainly Dell, Intel, IBM, Google, Microsoft and other multinationals present in Ireland can help the list grow further.

There are four decommissioned machines that would also rank on the feasibility list in terms of performance. It is envisaged that the Irish500 will also serve as a repository for systems as they are taken out of commission. This is in contrast to the other lists that stop tracking decommissioned machines as soon as they fail to rank. The Irish decommissioned machines (performance in Tflops) are: ICHEC [Schrodinger (11.1), Lanczos (4.7), Walton (3.14)] and TCD [IITAC (2.7)]. As the number of decommissioned machines grows, their data, combined with machines still in use will together help create a broader, more historical picture of the Irish HPC landscape as it evolves.

Finally, it is encouraging looking at the Top500 feasibility list coupled with the fact that the threshold of Top500-class is ~75Tflops, and that many machines on the Irish list above are very close to this. It is also interesting to see that even though this list which will certainly be added to, the top eleven Irish installations sum to a healthy 280 Tflops. Additionally, comparing this list to the current TSI list is encouraging. The Irish feasibility list has 11 systems compared to 27 for TSI and both the highest and lowest ranking systems on each list are comparable in performance (well within small multiples).

4.4 Irish500.org

The Irish500 website (www.irish500.org) will primarily function as a forum to present results to the public and participants but will also accept submissions. All submissions will be followed up by personal contact from the committee. The information collected from participants will include, but is not limited to: location, segment, system model, manufacturer, Rmax, power consumption, total cores, accelerator/GPU (cores/type/family), processor (generation/speed), RAM (per core/total), total system storage, OS, OS family, interconnect (speed/family), application area(s), total purchase cost, running costs and average utilisation.

4.5 Call for Participation

The Irish500 Advisory Committee has issued a public call for participation, located on the website, targeted at Irish HPC installation owners/organisers from all sectors. The Advisory Committee is also actively working on engaging these sectors to discover and confirm existing HPC installations. Later this year, the first Irish500 list will be released. Participation will occur both through the Irish500.org website and personal contact by the committee. The committee is also keen to expand by taking on new members, particularly from other institutions and private industry. We envisage a twice-annual release schedule similar to Top500, in June and November of each year. Within one year of the initial (performance-based) list, we envisage the first release of energy efficiency rankings.

4.6 Privacy and Anonymity

The Irish500 is committed to improving the public awareness of HPC domestically and internationally, and therefore all submissions of Irish-based installations, regardless of sector, are welcome. Due to the small size of the Irish market, the committee does recognise the possibility that, particularly in private industry, some companies may be reluctant to release certain information to the public for competition reasons. The Irish500 therefore provides for anonymous submissions, where only performance data and limited (agreed) hardware specifications are released to the public. The committee will need to verify other details of the installation. However these will not be released at the request of the installation owner.

5 Conclusion

This paper presents a case for an Irish500 list, ranking the top HPC installations geographically situated in Ireland by performance, and in the near future energy efficiency. The Irish500 project is motivated by two considerations. Firstly, it seeks to be a vehicle for the goals of advancing, advocating and promoting Irish HPC domestically and internationally. Secondly, Ireland has a unique position in the global HPC arena and there is a lack of a central and independent body to help achieve these goals. In particular the Irish500 seeks to answer the questions:

- *What is the landscape of HPC in Ireland today?*
- *Where does this landscape fit globally?*

The proliferation of the Top500 list format and the importance and success of other lists, justify adopting a similar format for the Irish500 list. Evidence that such lists have significantly impacted stakeholders in HPC, from practitioners to the general public further strengthens the case for the Irish500. A feasibility list is presented, *made up entirely of publicly-available information* describing eleven systems with a performance profile similar to the Top Supercomputers India list, and containing systems of global importance. Some of these systems are formerly ranked on the Top500 list, and every major HPC sector in Ireland is represented. This demonstrates that the Irish500 list could quickly be as successful as the Top Supercomputers India list (the only other known geographically-linked list), both in system numbers and performance. This would be a positive development for the Irish HPC community, a major step towards addressing the two questions posed above, and a significant factor in moving the Irish HPC community forward. We have formed an advisory committee and developed a prototype website. Following further expansion of the committee and direct contact between the committee and identified Irish HPC owners and maintainers, we will issue a general call for participation. After this, the first Irish500 list will be announced. It is intended that the Irish500 will be continuously open to submission, constantly updated and refined, and released on a bi-annual basis.

Acknowledgements

The authors would like to thank Catherine Mooney, Kevin Hely, Liam Power and Justin Smyth for useful assistance and advice.

References

[1] A. Apon, S. Ahalt, V. Dantuluri, C. Gurdgiev, M. Limayem, L. Ngo and M. Stealey, "High Performance Computing Instrumentation and Research Productivity in U.S.," *Journal of Information Technology Impact,* vol. 10, no. 2, pp. 87-98, 2010.

[2] "Top500 Supercomputer Sites - Project," [Online]. Available: http://www.top500.org/project/. [Accessed 20 February 2013].

[3] R. Smith, "Cineca's Tesla K20-Based "Eurora" Supercomputer Unveiled; Water Cooling Unlocks Extra Efficiency," 31 January 2013. [Online]. Available: http://www.anandtech.com/show/6717/cinecas-tesla-k20based-eurora-supercomputer-unveiled-water-cooling-unlocks-extra-efficiency. [Accessed 17 February 2013].

[4] J. Markoff, "The iPad in Your Hand: As Fast as a Supercomputer of Yore," *New York Times,* 9 May 2011.

[5] D. G. Feitelson, "The supercomputer industry in light of the Top500 data," *Computing in science & engineering* , vol. 7, no. 1, pp. 42-47, 2005.

[6] S. Sharma, C.-H. Hsu and W.-c. Feng, "Making a case for a green500 list," in *Proceedings of the 20th International IEEE Parallel and Distributed Processing Symposium,* 2006.

[7] A. Petitet, R. C. Whaley, J. Dongarra and A. Cleary, "HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers, Version 2.0," Innovative Computing Laboratory, University of Tennessee, 10 September 2008. [Online]. Available: http://www.netlib.org/benchmark/hpl/. [Accessed 17 February 2013].

[8] S. Vadhiyar, "Top Supercomputers India," Supercomputer Education and Research Centre, Indian Institute of Science, 27 December 2012. [Online]. Available: http://topsupercomputers-india.iisc.ernet.in/. [Accessed 17 February 2013].

[9] J. Barr, "Next Generation Cluster Computing on Amazon EC2 - The CC2 Instance Type,"

Amazon Web Services, 14 November 2011. [Online]. Available: http://aws.typepad.com/aws/2011/11/next-generation-cluster-computing-on-amazon-ec2-the-cc2-instance-type.html. [Accessed 25 February 2013].

[10] J. Barr, "High Performance Computing Heads East - EC2 CC2.8XL Instances in EU West (Ireland)," Amazon Web Services, 18 June 2012. [Online]. Available: http://aws.typepad.com/aws/2012/06/high-performance-computing-heads-east-.html. [Accessed 17 February 2013].

[11] "Cycle Computing Blog," Cycle Computing LLC, 29 September 2011. [Online]. Available: http://blog.cyclecomputing.com/2011/09/new-cyclecloud-cluster-is-a-triple-threat-30000-cores-massive-spot-instances-grill-chef-monitoring-g.html. [Accessed 17 February 2013].

[12] "Cycle Computing Blog," Cycle Computing LLC, 19 April 2012. [Online]. Available: http://blog.cyclecomputing.com/2012/04/cyclecloud-50000-core-utility-supercomputing.html. [Accessed 17 February 2013].

[13] H. Liu, "Amazon Data Center Size," 13 March 2012. [Online]. Available: http://huanliu.wordpress.com/2012/03/13/amazon-data-center-size/. [Accessed 17 February 2013].

[14] T. Trader, "Utility Supercomputing Heats Up," Tabor Communications, 28 February 2013. [Online]. Available: http://www.hpcinthecloud.com/hpccloud/2013-02-28/utility_supercomputing_heats_up.html. [Accessed 2 March 2013].

[15] "Grid-Ireland Closure on 31 December 2012," Grid-Ireland, 24 October 2012. [Online]. Available: http://www.grid.ie/closure.html. [Accessed 17 February 2013].

[16] R. C. Murphy, K. B. Wheeler, B. W. Barrett and J. Ang, "Introducting the Graph 500," Cray User's Group (CUG), 2010.

[17] J. Napper and P. Bientinesi, "Can cloud computing reach the top500?," in *Proceedings of the ACM combined workshops on UnConventional high performance computing workshop plus memory access workshop* , 2009.

[18] "HPC Challenge," University of Tennessee Innovative Computing Laboratory, 17 February 2013. [Online]. Available: http://icl.cs.utk.edu/hpcc/. [Accessed 17 February 2013].

[19] E. Strohmaier, "Insights from the Top500," *Scientific Computing,* pp. 14-16, August 2012.

[20] T. Trader, "Stanford Lights Up One Million Sequoia Cores," Tabor Communications, 28 January 2013. [Online]. Available: http://www.hpcwire.com/hpcwire/2013-01-28/stanford_lights_up_one_million_sequoia_cores.html. [Accessed 17 February 2013].

[21] B. Subramaniam and W. C. Feng, "Understanding power measurement implications in the green500 list," in *Green Computing and Communications (GreenCom), IEEE/ACM Int'l Conference on & Int'l Conference on Cyber, Physical and Social Computing (CPSCom)* , 2010.

[22] E. Strohmaier, S. Horst and J. Dongarra, "Top 500 Supercomputer Sites, November 2012," Top500.org, 12 November 2012. [Online]. Available: http://www.top500.org/lists/2012/11/. [Accessed 17 February 2013].

[23] "Trinity Centre for High Performance Computing," Trinity College Dublin, 15 February 2013. [Online]. Available: http://www.tchpc.tcd.ie/resources/clusters. [Accessed 17 February 2013].

[24] "High-Performance Computing (HPC) Infrastructure," Irish Centre for High-End Computing, January 2012. [Online]. Available: http://www.ichec.ie/infrastructure/. [Accessed 17 February 2013].

[25] G. Wang and T. E. Ng, "The Impact of Virtualization on Network Performance of Amazon EC2 Data Center," in *INFOCOM, 2010 Proceedings IEEE*, 2010.

Session 3

Text and Data Mining

Communities and Crime: Creating a Data Mining Prediction Model

Identify Future Crimes given the Socioeconomic Background in an Area

John Ryan[1], Markus Hofmann[2]

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

Abstract
This research paper outlines the methods involved in creating a prediction model from the "Communities and Crime Dataset", in order to forecast future specific crimes in a community as a result of various defined economic factors from within that dataset. While the initial results indicate that the dataset that the model is being trained on is unable to predict specific crimes with a sufficient high degree of confidence it can, however, with a high degree of accuracy, state that the area is one of low or high crime using aggregated class labels. Thus, this research paper presents proof that prediction models can be used with a resultant high degree of accuracy by grouping the class labels as opposed to prediction models which treat each crime as a specific class label.

Keywords: Communities and Crime, Prediction Model, Linear Regression, Decision Tree

1 Introduction

Unemployment causes crime rates to increase [1]. Lochner et al [2] estimates that if US high school graduation rates had been 1 percentage point higher in 1990 then there would have been 100,000 fewer crimes, 400 fewer murders, 8000 fewer assaults and the state legislature would have saved $1.1 billion from preventing murders. In 1993, criminal activity cost the United States $450billion [3]. In 2006, the United States spent $168 billion on police and corrections [4]. Consequently, there are real tangible social and financial factors that benefit from controlling and lessening crime in an area. For instance, as well as reducing the cost to the US legislature, a decrease in crime within a community results in an increase in local house prices from between 1% to 14.3% [5]. Datasets such as the "Communities and Crime Dataset" [8] provide opportunities for a locality's socioeconomic factors to be analysed with regards to the types of crime that may occur in that community and potentially construct a prediction model around them.

The dataset itself is comprised of 3 separate sources (1990 US Census, 1995 US FBI Uniform Crime Report, 1990 US Law Enforcement Management and Administrative Statistics Survey). There are some issues with the combined dataset as a result of this merging. For instance, there is an over reliance on data from some US States as opposed to others; 279 communities were surveyed within California and just 46 were surveyed in New York. Another example of this is the disproportion of ethnic populations; for example 1,137 communities had 90% White Caucasians while just 3 had a 90% African American majority. Missing values are a significant factor in the dataset sampling with 98 attributes having at least one missing piece of data. 28 of those 98 attributes have over 40% of their data values missing. There are 2,215 examples within the dataset and 147 attributes. Two Attributes are polynomial while the remaining 145 are numeric. The polynomial attributes (Community Name and State) do not have any role in the prediction. Out of the 147 attributes there are 18 potential class labels in this dataset (Table 1). The dataset's attributes are available from [8].

The prediction model results within Section 2.8 indicate that specific crimes cannot be specifically forecast given socioeconomic factors with a high degree of confidence. However, by combining the crime types as in Sections 2.9 and 2.10, other prediction models can forecast whether an area has the potential of being a high or low crime area with significant accuracy.

2 Methodology

2.1 Prediction Model

The Prediction Model for predicting the likelihood of specific crimes that is used for training and implementation is a Linear Regression algorithm model and was built using the CRISP-DM methodology. Linear Regression is used due to the class labels being numeric, and the dataset sampling itself consisting entirely of numeric values. It was chosen over the likes of Support Vector Machine (SVM) algorithms, as it can provide clear decisive performance parameters along with influential attributes on which crime factors the dataset sample is best at predicting. The values in question are the root relative squared error indicating the variance of the class label, so for example a variance for burglaries label at .33 (33%) would be much better than .77 (77%) for a murder class. Other beneficial factors of using Linear Regression are that the most influential attributes can be viewed and then sorted by their coefficient values from largest to smallest or vice-versa. Finally with the returned t-Stat and p-Value information, attributes can be eliminated that are below a standard. In this case attributes are considered valuable if they have a t-Stat >2 and p-Value (probability) <0.05, meaning that the attribute is at least 2 standard errors from zero and is 95% confident that the value is not zero. Attributes with high t-Stat values are more influential in the prediction model than ones with lower t-Stat values.

2.2 Prediction Model Implementation

The techniques of removing useless attributes and filter examples (exclude missing value attributes) proved to be of no benefit to the dataset sampling and were therefore excluded from the final prediction model. As stated in the introduction, one of the most significant issues was the amount of missing data values which had to be resolved first (Section 2.3). Noise and Outliers are examined in Section 2.4. Correlation was also a significant factor in order to explore the reduction in the number of attributes from the 147 present in the dataset (see Section 2.5). Finally, it may be worth exploring if a subset of attributes has potentially better accuracy than the original attribute subset. Two feature subset methods were used to generate these subsets; Forward Selection and Genetic Algorithms (see Section 2.6). The default class label used for binary classifications, unless otherwise stated, is "burglaries: number of burglaries in 1995".

2.3 Missing Values

A range of missing value substitution techniques were applied to the dataset, including methods such as Average Mean, Minimum, Maximum, Imputation (SVM) and Imputation (k-NN). Average Mean and Imputation (SVM) proved to be the best missing value techniques as they did not change the pattern of the dataset, which is critical [6]. The data value ranges and the standard deviation results for each of the attributes with substituted values was not altered by using these two techniques thus they did not influence the pattern in a different manner than before the dataset contained substituted values. To decide which of the two techniques would perform best on the dataset, a range of tests were applied to the dataset using both substitution methods. The tests included:

- Linear Regression with Listwise Deletion (to see if the 28 attributes that had over 40% of their data values missing could be removed) .
- Linear Regression without Listwise Deletion.
- Classification using binning and the k-NN algorithm with Listwise Deletion then without Listwise Deletion.

For Linear Regression a combination of factors decided the results including the variance (which of the two substitution methods produced a lower variance on the class label) and the number of influential attributes they produced with their t-Stat > 2 and p-Value <0.05. As a further reassurance measure, binning and the k-NN algorithm were used to obtain the performance accuracy measurements for both substitution techniques, again with and without Listwise Deletion. These performance results acted as a guide, so confidence in the selected substitution technique would be heightened if both the variance was low and the performance accuracy was high. The results from these tests (variance: Relative Root Squared Error .303 for Imputation and .326 for Average Mean) proved that Imputation using the SVM algorithm was the best missing value substitution technique. This is then combined with Listwise Deletion for an optimal performance, as using it returned favourable variance (variance: Relative Root Squared Error .309 for Imputation and .336 for Average Mean) and significant t-Stat and p-Value attributes. Although the variance is down slightly in Imputation (SVM) using Listwise Deletion, it is a negligible .06, the overall values within the t-Stat>2 and p-Values<0.05 attributes are higher using Listwise Deletion than when not using it and the performance accuracy for Imputation(SVM) with Listwise Deletion was 99.82%.

Table 1 Class Labels for Dataset [8]

	Attribute and Description
1	murders: number of murders in 1995
2	murdPerPop: number of murders per 100K population
3	rapes: number of rapes in 1995
4	rapesPerPop: number of rapes per 100K population
5	robberies: number of robberies in 1995
6	robbbPerPop: number of robberies per 100K population
7	assaults: number of assaults in 1995
8	assaultPerPop: number of assaults per 100K population
9	burglaries: number of burglaries in 1995
10	burglPerPop: number of burglaries per 100K population
11	larcenies: number of larcenies in 1995
12	larcPerPop: number of larcenies per 100K population
13	autoTheft: number of auto thefts in 1995
14	autoTheftPerPop: number of auto thefts per 100K population
15	arsons: number of arsons in 1995
16	arsonsPerPop: number of arsons per 100K population
17	ViolentCrimesPerPop: total number of violent crimes per 100K population
18	nonViolPerPop: total number of non-violent crimes per 100K population

2.4 Noise/Outliers

The attribute racePctWhite is highly skewed to the right in that it generates a left long tail normal distribution graph which indicates the majority of its values are above their average mean. The other ethnic groups in the dataset have the majority of their values below their average mean, with right long tail normal distribution graphs. Thus there could be bias and this should be noted in the final results. These values are more than likely genuine figures and not as a result of other external factors such as mistyping. The ranges of values for each attribute within metadata analysis showed no obvious exceptions. Missing value substitution could not cause the skewness as there were very few missing values present. A possible reason for the rate of bias can be understood in that 1,137 of the surveyed communities had majority population rates of White Caucasians (over 90%), no Asian group managed higher than 57% (that was in California) and 11 communities had 90% Hispanic (the two largest communities are based in Texas). When this type of information is compared with the same communities from the US Census 2010 [10] the data correlates favourably with the ethnic population portions (Table 2).

A Scatterplot (Figure 1) demonstrates that the data points are highly clustered and that 3 of the outliers are actually within that cluster, or just on the border edge of them. So this may mean that they are legitimate values, but just at the extreme end of them. Consequently they are genuine anomalies. There are definitely 2 distinct outliers, as the remaining 3 are just outside the cluster but are alongside 3 legitimate data points.

Talend Open Profiler was used to check for issues such as duplicate records. It ruled out duplication between community name and US State as it showed that there are multiple communities with the same names based in separate US States, such as Jacksonville City. In summary, after analysing the data through the software, there did not seem to be any noise that was evident enough to be an issue with the final dataset.

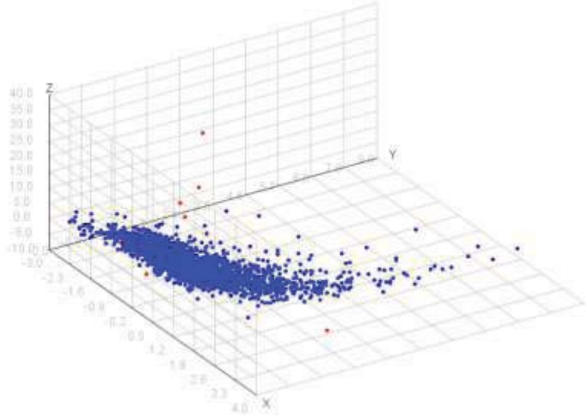


Figure 1 Distance Outlier Detection k=10

Table 2 Comparing Ethnic Proportion between US Census 2010 and Data Sample

State	Comparing Communities between the US Census Data 2010 and the Communities & Crime Dataset			
	Community	*Ethnicity*	*US Census 2010*	*Communities & Crime Dataset*
Indiana	Jasper City	White Caucasians	93.6%	99.63%
California	Monterey Park City	Asian	66.9%	57.46%
Texas	Eagle Pass City	Latino & Hispanic origins	95.5%	95.29%

2.5 Correlation

A total of 18 out of the 147 attributes are class labels so 17 can be excluded from each analysis when using Linear Regression. Listwise Deletion will be included in the final model so a further 28 attributes are removed, leaving the dataset at 102 attributes with 2,215 examples. Or 100 when the 2 polynomial attributes are removed, as they add nothing in terms of predictive power. Thus, it is worth exploring if the number of attributes could be further reduced, so correlation between the attributes was analysed and taken into consideration during this pre-processing stage. Correlation displays highly correlated attributes that are as close to -1 or +1 depending on whether they are negatively or positively correlated. The attributes that were deemed highly correlated made sense subjectively. For example the attribute “PctWorkMomYoungKids: percentage of moms of kids 6 and under in labor force” is highly correlated with the attribute “PctWorkMom: percentage of moms of kids under 18 in labor force" with a correlation value of 0.904. Another example would be the attribute “PersPerOccupHous: mean persons per household" being highly correlated to attribute “PersPerFam: mean number of people per family" which has a correlation value of 0.933. In summary, 36 attributes were deemed to be highly correlated and could potentially be removed from the dataset for the final prediction model.

2.6 Weighting Attributes- Feature Subsets

Two Feature Subset methods were used to create a subset of attributes that would reduce the number of attributes without the model reducing its prediction accuracy; these were Forward Selection (using the ranking method Weight by Information Gain and then select the top 60 attributes by Weights) and Genetic Algorithm. For Forward Selection, two tests were implemented on the dataset with both correlation attributes included and then excluded to measure the performance. Decision Trees, k-NN and Naïve Bayes using Binning were explored. The optimal performance was not having any correlated attributes and having a bin size of 1200. All three algorithms produced the same attribute subset of 6 attributes, but with different levels of accuracy; Decision Trees was 81.81%, Naïve Bayes 74.27% and k-NN 83.34% accurate. Therefore k-NN was deemed to be the most accurate and its subset was selected. As stated, all 3 algorithms produced the same 6 attribute subset, which was:

1. NumKidsBornNeverMar: number of kids born to never married
2. population: population for community
3. PctYoungKids2Par: percent of kids 4 and under in two parent households
4. PctRecImmig10: percent of _population_ who have immigrated within the last 10 years
5. racepctblack: percentage of population that is African American
6. PctHousOwnOcc: percent of households owner occupied

In conclusion, after analysing the data, it is evident that the dataset sample required all its attributes to produce an accurate subset of attributes for a better performance. Forward Selection using Weight by Information Gain returned just the 6 attributes subset which may not be enough socioeconomic data to accurately create a pattern for the final prediction model. Thus it should be compared with another Feature Subset method for a further examination of what the best subset of attributes would be. In this case the Feature Subset method, Genetic Algorithm, was used as it can handle a large number of attributes. This included removing the 36 correlated attributes and using the Linear Regression algorithm. This returned with a variance of 0.328 and a 40 attribute subset, deeming them to be the best attributes of interest. The selection schema was the Roulette Wheel with a maximum fitness score of 10.

2.7 Selecting the Dataset for the Final Prediction Model

The original dataset is slightly modified at this stage with the 28 attributes removed through Listwise Deletion, the 36 attributes excluded due to high correlation and the replacing of missing values using Imputation (SVM). Both Forward Selection and Genetic Algorithm have produced subsets of attributes that may be better than this modified original subset. In order to discover which set of attributes is included in the final model, a Linear Regression algorithm is executed using two class labels on the 2 attribute subsets and the original modified dataset sample:

1. burglaries: number of burglaries in 1995
2. autoTheft: number of auto thefts in 1995

For the Burglaries class label the variance for the original dataset is far lower at .305 than the Genetic Algorithm attribute subset which is .352 and the Forward Selection subset at 0.323. For Autotheft the Genetic Algorithm attribute subset has a much worse variance at .366 compared to the original subset at .326 and Forward Selection attribute subset at 0.332. The original subset for Burglaries class returned more influential attributes (t-stat>2 and p-Value<0.05) than the Genetic Algorithm attribute subset managed, ratio 12 versus 9. The Genetic Algorithm subset for Autotheft class returned marginally more influential attributes (t-stat>2 and p-Value<0.05) than the original dataset managed, ratio of 8 to 5. Although the number of attributes and the overall performance results indicate that the original dataset should be used, the Genetic Algorithm subset of attributes is processed through a final test by executing a Linear Regression algorithm for class label Robberies using both the original dataset and the Genetic Algorithm subset. The original dataset returned a variance of .383 whereas the Genetic Algorithm returned .418, thus ruling it out completely.

Normalisation was applied to both class labels using the same Linear Regression algorithm to smooth out data points to ensure that anomalies are not causing issues with the dataset results. The variance returned was almost identical to the non-normalised values, for burglaries the original was .305 and for autotheft the value was .325 for the original dataset. Thus it adds nothing to the final model, and proves that the data is consistent throughout the dataset sampling so the missing value substitution technique did not affect the pattern. It can be conclusively written that normalisation will not be part of the final model.

In summary, the original dataset with the modifications previously outlined appears to be the better option than the subset of attributes produced from either the Genetic Algorithm or Forward Selection. Thus the dataset used in the final prediction model will be the original dataset with the following amendments:

- 17 class labels removed with just the one used out of the 18.
- 28 attributes removed that had 40% of their values missing.
- 36 attributes removed as they were deemed and proven to be highly correlated and to be of benefit to the model’s performance by being removed.
- Imputation (SVM) being used to replace the missing values in the dataset sampling.
- 2 polynomial attributes removed that add nothing to the model.

2.8 Prediction Model Results: Identifying Specific Crimes in an Area

The results for burglaries and autothefts were known from previous research work using the modified original dataset. The prediction model underwent a comprehensive series of Linear Regression tests on other class labels to judge the accuracy of the final model by discovering what their variance was. The class labels were grouped by their diverse missing values to obtain an overall performance perspective. “Original” is the original dataset that just has missing value substitution using Imputation (SVM) and “Modified” is the new modified dataset. These two columns are compared to see if the new dataset provides a better performance. As can be observed in Table 3, the overall majority of the results show an improvement in the variance between the original and the modified datasets. However, there is no set pattern to the prediction model. For example, if there was a low variance for classes with few missing values then it could be assumed that the prediction model works for the most complete attributes, but it cannot be that way. The Robberies class, with more missing values than AutoTheftPerPop, has a better root relative squared error. Thus, the dataset sampling cannot be a good predictor of crime using these surveyed socioeconomic factors because the performance is poor.

The class labels robberies, burglaries and larcenies were merged to discover if combining class labels together may help the performance. These three classes were aggregated as they could all be considered under the same category of theft. When this new class label was processed through a Linear Regression algorithm the variance was .330 +/- 0.114, and the influential attributes returned were similar to previous class tests, in this case “ PctNotSpeakEnglWell: percent of people who do not speak English well” and “ NumKidsBornNeverMar: number of kids born to never married”. Overall 23 attributes were returned with a t-Stat >2 and p-Value < 0.05.

There are other types of prediction models, aside from the Linear Regression Model, that can provide improved ways of evaluating the prediction quality of this dataset once the class label is changed from solely predicting specific crimes such as Burglaries and Autotheft. For instance, by creating a new binomial class label entitled “CrimeClass” with just 2 classes High Crime and Low Crime (Section 2.9) then algorithms such as Decision Tree and k-NN can be used to predict to a very high degree of accuracy whether there is a High Crime Rate or Low Crime Rate within a community depending on its localised socioeconomic factors.

Table 3 Prediction Model Results using Linear Regression

Class Label	**Class Labels Variance between Original Dataset and Modified Dataset**		
	Missing	**Original**	**Modified**
Murders	2016	0.820	0.757
Robberies	796	0.495	0.383
Assaults	316	0.674	0.563
larcenies	3	0.355	0.372
autoTheftPerPop	3	0.668	0.698
Arsons	1471	0.806	0.789
nonViolPerPop	97	0.733	0.708

2.9 Prediction Model Results: Binomial Class Label CrimeClass: Low and High Crime

Another approach is creating a binomial class “CrimeClass”, with class labels High Crime/Low Crime, by adding 8 of the main crimes (Reference Rows 1, 3, 5, 7, 9, 11, 13 and 15 in Table 1) and splitting the crimes between a high class and a low class depending on whether they are larger or smaller than the average mean (4408.4[1]) that was generated from the total addition of

[1] Notes on 4408.4: Adding the 8 Main Class Labels within Excel 2010 that represent the total number of crime category amounts and labelling it as “OverallCrime”. Obtaining the mean of “OverallCrime” over the 2,215 rows, which is 4,408.4. Creating a new column “CrimeClass” with the following formula to populate it, =IF(EN2216<$EN$2217, "LOW", "HIGH"), with $EN$2217 representing the overall average number 4,408.4 and EN2216 representing the cell for the “OverallCrime” value per row. “CrimeClass” identifies Low Crime as less than 4408.4 and High Crime as greater than or equal to 4,408.4. This creates a class

the 8 crimes. With this sample, there is a class imbalance between Low and High Crime in the ratio of 5.45:1 with Low in the majority. In all there are 1872 communities with a low crime rating and 343 communities with a high crime rating. This binomial class label allows for the use of algorithms such as Decision Trees. Using Information Gain as the criterion with the default settings of the Data Mining software RapidMiner (Minimal size for split=4, Minimal leaf size=2, Minimal gain=0.1, Maximal depth=20, Confidence=0.25, number of validations=10), thus with no configuration changes to aid performance, provides an accuracy of 94.58%. The Decision Tree contains significant information about how the attributes (the socioeconomic factors) link in a logical flow and also shows that the main root node is the number of people living in the urban community, which is then split between poverty and the number of children born to unmarried parents. Unfortunately the Decision Tree is too large to be represented properly in this research paper.

The algorithm considers the main split from the root node if the number of people classified as urban is greater than 66,366 then the percentage of people under the poverty line needs to be considered, otherwise the number of kids born to unmarried parents (attribute "NumKidsBornNeverMar..") is a factor. Interestingly, an attribute considered important is the number of vacant households in communities, it is considered relevant for areas that also have over 960 children with unmarried parents. Vacant households is an attribute, like number of kids to unmarried parents, that appears in numerous results within this research as a significant factor indicating that this is an attribute that does require consideration with crime. Also, as an aside, the number of kids born to unmarried parents is shown to be highly correlated with the number of vacant households.

The subtree that resulted from communities whose population portion under the poverty level is rated at less than (or equal to) 8.28%, clearly shows that there are numerous considerations with regards to crime and how it is interpreted within this subtree from the poverty level. The next node split from this is the percentage of the community that are divorced. If this figure is above 10.68% then age is the next consideration. If the percentage of 12-29 year olds in a community is above 29.34% then there is a high potential rate of crime and if less than 72.05% of the area has poor English language speakers there is a low prediction of crime. An item that should be noted about this subtree is the rate of potential misclassification for the prediction of a high rate of crime for those areas that have more than 74.72% of their population who are employed and who are over 16 years of age. The subtree that is generated from areas where the poverty level is greater than 8.28% register factors that indicate high degrees of crime:

- Areas that have less than 29.71% of people aged 16 or over employed in manufacturing.
- Communities where less than 29.92% of households have social security.
- Areas where less than 58.82% of the household units are fewer than 3 bedrooms.

A few tasks were attempted to increase the accuracy while applying 9-fold cross validation. In order to make the Decision Tree more flexible and less prone to potential overtraining, the Minimal Leaf Size was adjusted to 12 as too small may mean that there would only be a limited number of rows for matching and the Minimal Size for Split was set at 6, so it is not over trained (the default is 4). With all this, the accuracy has increased to 95.08%. If the bootstrapping sample size operator is added to this changed model then the accuracy increases to 96.77% but the genuine data pattern is diluted as the sampling of the records are now not all genuine survey data as there may be a disproportion of certain classes added [7]. k-NN is another algorithm that

label of LOW (1872) and HIGH (343) values. So there immediately seems to be a class imbalance that should be noted, with Low Crime being in the majority.

could be used. Using its RapidMiner software default settings (k=1, Measure Types=Mixed Measures, Mixed Measure=Mixed Euclidean Distance), the algorithm provides an accuracy of 92.73%. This could potentially be improved by changing variables such as the k value, which may be a worthwhile exercise as the ROC[2] curve in Figure 2 indicates that it may be more accurate than the Decision Tree with a fractionally larger area under its curve [7].

2.10 Prediction Model Results: Numeric Class Label : TotalCrime

Linear Regression is another Prediction Model that could be used, but this time for one single numeric class label entitled “TotalCrime”. As previously mentioned in this report, Linear Regression would facilitate the understanding of what attributes contribute most to the prediction model by returning measurable values such as t-Stat, p-Value and Standard Deviation Coefficients. The Root Relative Squared Error provides the level of variance in the prediction model. Creating a class label called “TotalCrime” adds 8 of the main crimes (Reference Rows 1, 3, 5, 7, 9, 11, 13 and 15 within Table 1) and uses this numeric class label within a Linear Regression algorithm. In this instance the variance returned is a Root Relative Squared Error of 0.279 +/- 0.079, which is much better than any of the variance obtained from the Linear Regression model when trying to predict specific crimes. The downside was that the result was computationally expensive. The attributes with t-stat>2 and p<0.05 register the number of children born to unmarried parents with a t-stat value of 37.25. Another significant attribute resulting from this process is the extremely high standard coefficient of the number of police per 100,000 people. This indicates a link between higher police numbers and lower crime rates. Finally the returned std. coefficient value of 5.30 for number of kids born to non-married parents (attribute "NumKidsBornNeverMar...”) supports findings throughout this paper, including what was demonstrated in Section 2.9.

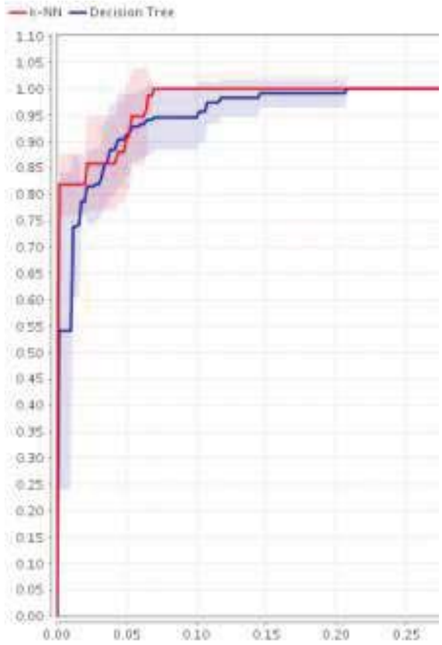


Figure 2 ROC Curve: K-NN (Top Line) and Decision Tree for Binomial Class Label: CrimeClass

[2] k-NN has a fractionally larger area under its curve (though there is a brief crossover of the lines at 0.5). This larger gap can be observed, for example, where the Decision Tree is closer to .010 than k-NN, though it is marginal at these default settings of k-NN with k=1 and Decision Tree using Information Gain, 4 as the minimum split, 2 as the minimum leaf size, 0.1 as the minimal gain and 20 as the maximum depth. The ROC sampling type is stratified sampling.

3 Conclusion

This research paper had a specific objective; whether the “Communities and Crime Dataset” sample could be used to create a prediction model which would identify potential specific crimes in a locality. While the prediction model was not sufficiently accurate enough to be considered as a useful tool in predicting specific crimes in an area given socioeconomic information (Section 2.8) it was successful in forecasting high and low crime within an area (Section 2.9), as well as providing valuable visual evidence of attribute relationships within Decision Trees (Section 2.9).

To create this model, the paper reviewed the different techniques required to create a final prediction model including handling missing values correctly and dimension reduction with correlation.

There is some outlying work that could be done to improve the prediction model:

- Improve the dataset by merging/adding similar datasets.
- Combining a much later dataset, perhaps from 2000 onwards (to reflect the reduction in crime rates indicated in the later 1990s [12]).
- Manually locate missing values and input them into the system.
- 14 of the 18 class labels have incomplete data. 71.43% of those have over 100 missing items each. 4 of those have over 40% of their data missing with the highest being the number of arsons (66.41%) and murders (91.02%). Merging datasets from other sources such as [9] may be of benefit.

Without these types of measures, a significant portion of the data is generated from computed data using Imputation (SVM) and not actual raw surveyed data.

This research paper has proven that an accurate prediction model can be constructed if the class label is adjusted. As previously stated, the original objective of specifying which exact crime cannot be resolved with a satisfactory degree of accuracy, but if the objective is changed to predicting high/low crimes then the prediction model can be modified to gain this improved accuracy. Thus, reducing the granularity of the prediction class allows the socioeconomic data to point to whether they cause a higher or lower risk of crime within a community but not whether they cause burglaries, for example. The best variance with the original class label and retaining the original objective was .305 (Section 2.8). By combining the 8 main crimes the class label could be changed to a binomial data type with a resultant accuracy of 94.58% (Section 2.9) or the class label could be kept as a numeric data type, which results in a much improved variance of .279 (Section 2.10).

The main attributes that kept registering in the prediction model process were also key factors reflected from existing socioeconomic research; namely factors such as education [13], age [14] and inequality [14]. The number of children born to unmarried parents was a particularly prominent attribute. There are existing programs in place to help these social issues, such as PALS in Canada [11], and the success of these are measurable. Thus, overall, the dataset sampling has proved to be a valuable source for identifying key trends that affect the criminal rate fluctuations in a community as its significant attributes produced do corroborate existing research.

In conclusion, the “Communities and Crime Dataset” contained valuable information that despite the potential bias and the definite issues of missing values, bore strong correlation to known and proven causes of crime rate fluctuations in localities. Thus ensuring that the prediction models created (such as the modified binomial class: “CrimeClass”) are relevant. As referenced in Section 2.8, the original prediction model requires its class labels to be modified to obtain the best quality outcomes and if not done, then more genuine data should be supplanted in order to investigate if that improves the outcome.

References

[1] D.T. Altindag. “Crime and unemployment: Evidence from Europe” International Review of Law and Economics, 32, 2012 pp. 145– 157.

[2] L. Lochner and E. Moretti. “The effect of education on crime: Evidence from prison inmates, arrests, and self-reports”. *The American Economic Review*, Vol. 94, Number 1, March 2004, pp. 155-189.

[3] B.C. Welsh and D. P. Farrington. “Monetary costs and benefits of crime prevention programs,” Crime and Justice, Vol. 27, 2000, pp. 305-361.

[4] S.N. Durlauf and D.S. Nagin. Overview of “Imprisonment and crime: can both be reduced?” - Executive Summary Imprisonment and Crime. *Criminology & Public Policy*, Vol. 10, Issue 1, 2011.

[5] D.G. Pope and J.C. Pope. “Crime and property values: evidence from the 1990s crime drop”. *Regional Science and Urban Economics 42*, 2012, pp. 177-188.

[6] D. Pyle. *Data Preparation for Data Mining*. Morgan Kaufmann Publishers, CA, USA, 2012. ISBN:1-55860-529-0

[7] J. Han, M. Kamber and J. Pei. *Data Mining Concepts and Techniques 3rd Edition*. Morgan Kaufmann Publishers (Imprint of Elsevier), MA, USA, 2012. ISBN: 978-0-12-381479-1

[8] M. Redmond (2011). *Machine Learning Repository - Communities and Crime Unnormalized Data Set* [online]. Accessed: 1 October 2012. Available: http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized.

[9] The Federal Bureau of Investigation (FBI) Uniform Crime Reports (2010). *FBI-Table 12 - Crime In the United States 2010* [online]. Accessed: 1 October 2012. Available : http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/tables/10tbl12.xls.

[10] U.S. Census Bureau (2010). *U.S. Census Bureau- State & County QuickFacts* [online]. Accessed: 1 October 2012. Available: http://quickfacts.census.gov/qfd/index.html.

[11] M. Cameron and C. MacDougall. "Crime prevention through sport and physical activity”. *Australian Institute of Criminology Trends and Issues in Crime and Criminal Justice*, 165, September 2000, pp. 1-6.

[12] G. LaFree. “Declining Violent Crime Rates in the 1990s: Predicting Crime Booms and Busts”. *Annual Review of Sociology*, Vol. 25, 1999, pp. 145-168.

[13] L. Lochner and E. Moretti. “The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports”. *The American Economic Review*, Vol. 94, No. 1, March 2004, pp. 155-189.

[14] Morgan Kelly. “Inequality and Crime”. *The Review of Economics and Statistics*, Vol. 82, No. 4, November 2000, pp. 530-539.

Speech Analysis of the 2012 American Presidential Election

Eugene Galvin [1], Markus Hofmann [2]

[1] Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Ireland
galvineugene@gmail.com

[2] Institute of Technology Blanchardstown, Blanchardstown Road North, Dublin 15, Ireland
markus.hofmann@itb.ie

Abstract

This paper shows the application of novel text mining techniques to analyse the fundamental differences of the campaign strategies between the Democratic and Republican parties by focusing on the campaign speeches of the 2012 American Presidential Election which pitted the incumbent Democrat Barack Obama against the Republican Mitt Romney. By analysing the speeches made by both men in the course of the campaign the strategies and government objectives of each side are be determined. With the use of text mining techniques such as text categorisation and similarity analysis techniques coupled with visual aids of the significant words and phrases, the fundamental differences between both parties become apparent.

1 Introduction

Every four years the American Presidential Election captures the imagination of not just the American public, but the entire global community. Such is the perceived influence of the elected leader of the 'Free World' in world economic, military and political issues, that in the year long run up to the election, blanket international media coverage is given to this divisive showdown between America's red and blue states. This practical paper analyses the fundamental differences between the Democratic and Republican parties by focusing on the campaigning strategies of the 2012 American Presidential Election which pitted the incumbent Democrat Barack Obama against the Republican Mitt Romney. By analysing the speeches made by both men in the course of the campaign the strategies and government objectives of each side are determined. This not just highlights the key issues of the campaign but by using similarity measures it denotes if either side is espousing a clear programme for government or is giving particular sound bites which appeal to a certain sector of the electorate. Dissimilarity measures emphasise the opposing views of each side, as to how to take the country forward over the next four years. These measures also determine whether the nominee acts reactively to his opponent or takes an independent proactive stance on set issues. With the use of text mining applications such as text categorisation and similarity analysis techniques coupled with visual aids of the significant words and phrases, the fundamental differences between both parties become apparent.

2 Data Collection

The data for this paper was initially obtained from two different sources. The speeches, which totalled twenty five per candidate, were extracted using various web crawling and web extraction processes through the RapidMiner [1] Mining tool. The primary data for both presidential nominees was collected from the *http://www.presidency.ucsb.edu/index.php* website. Additional data for Mitt Romney's speeches which were subsequently not used due to a large number of duplications were extracted from the *http://mittromneycentral.com/* website. Complications arose when trying to collect the data from the *http://www.presidency.ucsb.edu/index.php* website. A number of factors which may have contributed to this difficulty were the *url* structure combined with the number of speeches on the website. The website contains numerous amounts of previous election documents all of which were formatted with the same *url* structure as the specific speeches required. The collective size of these documents restricted the webcrawler's extraction ability even with adjusting the properties of the webcrawler to deal with the maximum number of webpages, the depth and the maximum page size. To overcome this a *.csv* file was created for each candidate which contained the *url* of each specific speech required.

3 Methodology

Some applications of text mining techniques to analyse aspects of political life can be found in [2], [3], [4] and [5].To determine if different strategies were adopted by both the Democrat and Republican nominees, the following analysis techniques were preformed:

- Speech Categorisation
- Speech Similarity
- Word Usage

3.1 Speech Categorisation

Initially to obtain an understanding of the different campaign approaches by both parties a categorisation task was preformed. This process also determined if it were possible to categorise additional unseen speeches by each candidate. A text categorisation model was generated. This technique can also be used to provide a word list detailing the frequency each candidate used various words (or tokens) in their campaign. Table 1 shows the RapidMiner operators and their paramter settings used to build and fit the classication model. Given the fundamental differences of the two candidates it was anticipated that a very high accuracy level can be achieved.

Table 1: Categorisation Model

Operator	**Parameter**	**Additional Comments**
Word Vector	TF-IDF	no pruning
Extract Information	Xpath	*//h:*[@id="content"*]
Tokenise	Non leters	
Transform Case	Lower case	
Filter Stopwords	RapidMiner list	
Stemmer	Porters	
Remove Duplicates	Attribute filter type – all	
X-Validation	# of validations -10	
Naive Bayes		

3.2 Speech Similarity

To establish if a clear government programme and campaign approach was adopted, similarity techniques were generated to determine if any comparisons can be drawn from the speeches. These techniques were also explored by [6] and [7]. Speeches which are similar to each other would indicate a more clear and consistent campaign approach, whereas speeches which are dissimilar would point towards an inconsistent and ad-hoc campaign process. Three similarity analysis techniques were generated; the similarity of Barack Obama's speeches, the similarity of Mitt Romney's speeches and a similarity measure of both nominees' speeches. The final analysis provides more insight to determining if both parties placed similar emphasis on the same topics. All similarity models were generated using the same operators as can be seen in Table 2 These models consisted of extracting the speeches from the website using different *.csv* files for each party. A combined *.csv* file was created to determine the similarity of both candidates speeches together with text files *1-26* denoting Obama's speeches while text files *27-52* signifying Romney's speeches. Given the general focus of identifying similarities we have only used the Cosine Similarity measure which measures the cosine of the angle between two vectors of an inner product space.

Table 2: Similarity Measure Model

Operator	**Parameter**	**Additional Comments**
Read .csv		
Get Pages	Link Attribute – Link	
Data to Documents		
Process Documents	TF-IDF	No pruning
Data to Similarity	Measurement type Numerical Measure	Numerical Measures CosineSimilarity
Tokenise	Non letters	
Transform Case	Lower case	
Stopwords List	RapidMiner list	

3.3 Word Usage

To ascertain which words were most used and how they were used a number of techniques were applied. Initially both sets of documents were loaded and pre-processing was applied as can be seen in Table 3. The resulting word count was analysed in Rapidiminer. Additional word processing was completed for both parties by using a linguistic tokeniser in RapidMiner to extract all sentences within each candidate's speeches and then analysing these sentences using the visualisation tools provided by *Many Eyes* (*http://www-958.ibm.com/software/analytics/manyeyes/)* such as word trees, tag clouds and phrase nets. All of these methods provided interesting insight into the strategy of each campaign.

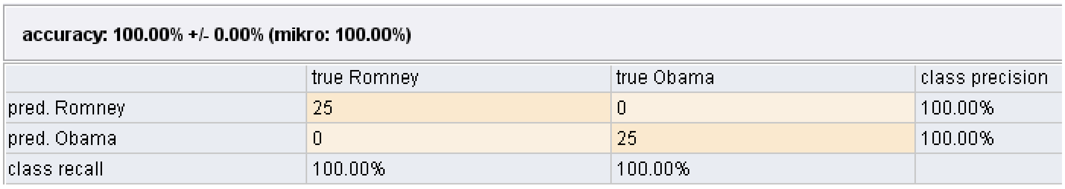
Table 3: Similarity Measure Model

Operator	**Parameter**	**Additional Comments**
Read .csv		
Process Documents	TF-IDF	No pruning
Extract Information	//h:*[@id="content"	
Tokenise	Non letters / Linguistic Sentences	
Transform Case	Lower case	
Stopwords List	RapidMiner list	
Filter Tokens	Min Chars- 4 Max Chars - 25	

4 Results

4.1 Speech Categorisation

The categorisation results shows a clear division between both candidates. All speeches were correctly classified with both categories achieving 100% accuracy (in all 10 iterations of the model training and testing phases) as can be seen in Figure 1. If only one speech was classified incorrectly it may provide an indication that similar techniques and approaches were being adopted by both sides, however it is reasonable to state that speeches from both candidates can be clearly identified. This model can also be used to categorise additional unseen speeches made by both people.



accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true Romney	true Obama	class precision
pred. Romney	25	0	100.00%
pred. Obama	0	25	100.00%
class recall	100.00%	100.00%	

Figure 1: Categorisation results in form of a Confusion Matrix

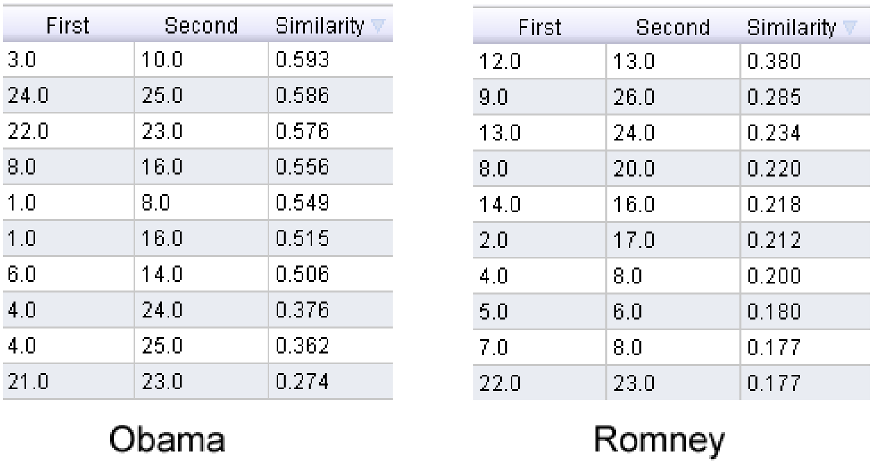
4.2 Speech Similarity

Analysis of speech similarities was conducted both individually and combined.

There is clear evidence of strong similarities within President Obama's speeches as can be seen in Figure 2. Seven similarities of over 50% were detected in the analysis and a further thirteen similarities of over 20% were clearly distinguished. The highest similarity between Obama's speeches was 59%. This provides additional evidence that a clear campaign programme was adopted by the Democratic Party in the 2012 election.

Speeches presented by Mitt Romney were not as cohesive as his opponents. The highest similarity between the speeches made is only 38% as seen in Figure 2 with only seven speeches achieving over 20% similarity. This indicates that the contents and therefore the topics addressed within the speeches were inconsistent and points to a more ad-hoc campaign approach.

An analysis of all speeches combined shows that there is little similarity between the speeches made by both men. None of the top twenty similarities were between speeches presented by both rivals.



First	Second	Similarity ▽
3.0	10.0	0.593
24.0	25.0	0.586
22.0	23.0	0.576
8.0	16.0	0.556
1.0	8.0	0.549
1.0	16.0	0.515
6.0	14.0	0.506
4.0	24.0	0.376
4.0	25.0	0.362
21.0	23.0	0.274

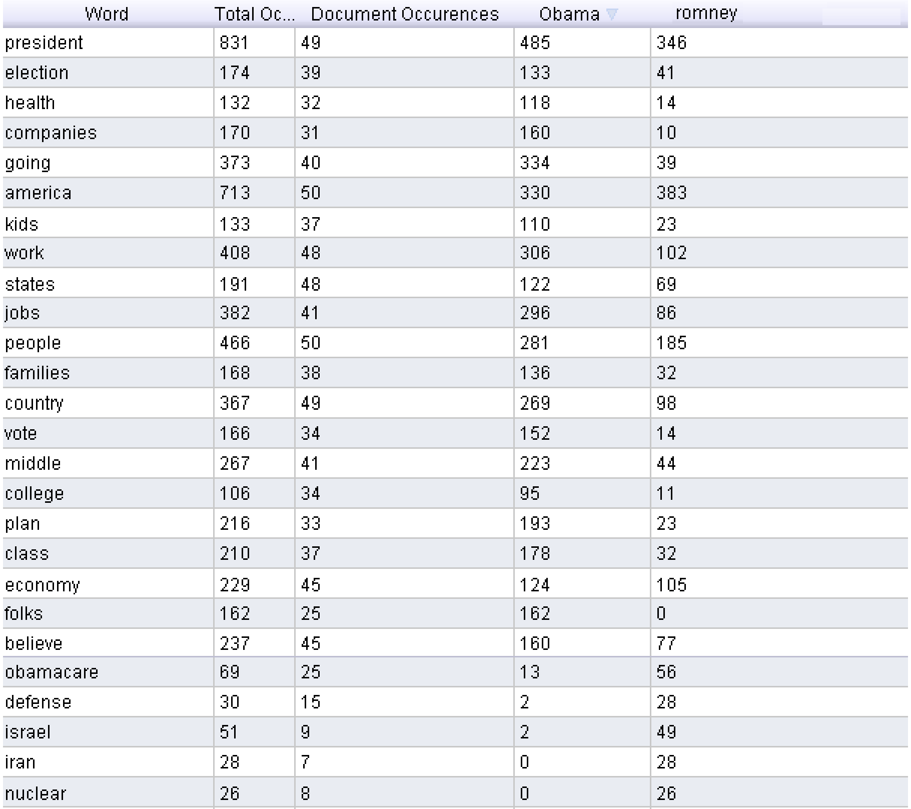
Obama

First	Second	Similarity ▽
12.0	13.0	0.380
9.0	26.0	0.285
13.0	24.0	0.234
8.0	20.0	0.220
14.0	16.0	0.218
2.0	17.0	0.212
4.0	8.0	0.200
5.0	6.0	0.180
7.0	8.0	0.177
22.0	23.0	0.177

Romney

Figure 2: Overall Similarity measures in descending order - top 10

In addition to the interesting findings with the similarity analysis other indication as to word usage can be seen. A selection of some of the words used by both candidates can be seen in Figure 3. It can be deduced from this list that Obama concentrated on the domestic issues within his speeches as indicated with the high usage of words such as *health*, *companies*, *jobs*, *families* and *college*.Mitt Romney's campaign strategy focused on discrediting his opponent, attacking Obama's term in office and the initiatives such as his health programme. Romney used the word *Obamacare* in an attempt to associate negativity with Obama and his policies. Additionally words such as *Israel*, *Iran*, *nuclear* and *defence* were more prevalent in Romney's speeches where as only on four occasions Obama used two of these words. The words *Iran* and *nuclear* were not mentioned in any of President Obama's speeches. This provides evidence that a branch of Romney's strategy was to appeal to foreign policy issues; which is further clarified by the locations where some of his speeches were presented. While all of Obamas speeches were presented nationally, Romney addressed audiences in Europe and the Middle East, highlighting the existing and potential threat of foreign forces to Americas domestic security.



Word	Total Oc...	Document Occurences	Obama ▽	romney
president	831	49	485	346
election	174	39	133	41
health	132	32	118	14
companies	170	31	160	10
going	373	40	334	39
america	713	50	330	383
kids	133	37	110	23
work	408	48	306	102
states	191	48	122	69
jobs	382	41	296	86
people	466	50	281	185
families	168	38	136	32
country	367	49	269	98
vote	166	34	152	14
middle	267	41	223	44
college	106	34	95	11
plan	216	33	193	23
class	210	37	178	32
economy	229	45	124	105
folks	162	25	162	0
believe	237	45	160	77
obamacare	69	25	13	56
defense	30	15	2	28
israel	51	9	2	49
iran	28	7	0	28
nuclear	26	8	0	26

Figure 3: Word usage in terms of occurrence (descending order of Obama)

4.3 Word Usage

As described in Section 3.3 numerous word usage analysis was performed using the many eyes visualisation tools. Initially speeches presented by Barack Obama were assessed followed by Mitt Romney's speeches.

Barack Obama

The initial analysis of these speeches was of word frequency by using the *Tag Cloud, word tree and phrase net* tools. These visual tools highlight instantly the topics and issues that were foremost in the Obama campaign as seen in Figures 3 through to Figures 5. Appealing to the American middle classes was a key plan in the Obama strategy. This is evident with the high word count of *middle* and *class* as seen in Figure 3 and the noticeable presence of the term *middle class* in Figure 4. He highlighted that large multinationals were busy exporting jobs abroad and that he was committed to keeping jobs at home. His focus on the Automobile Industry in the Midwest State of Ohio, by pumping millions of federal finances into it so that it could remain in America, which he constantly referred to in his campaign. This can be seen Figure 4 where a regular occurrence of the words *auto industry* is apparent.

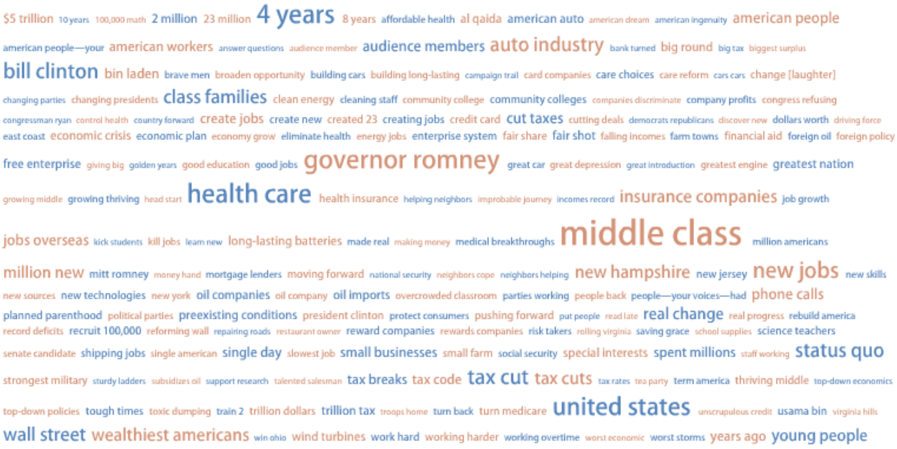


Figure 4: Tag cloud of Obama’s campaign Speeches

He also linked Bill Clinton to his campaign, indeed using the elder statesman during the campaign itself. In his speeches he harked back to the Clinton era and how 23 million jobs were created under this democratic administration. Obama talked about saving and fighting for the middle classes, he highlighted the positives of Obamacare but did not dwell on it. He appealed to the voters to vote for him so that he could fight for them in congress, cut their taxes, keep jobs at home, and continue with building the economy. Obama’s strategy was clear, concise and kept to domestic issues.

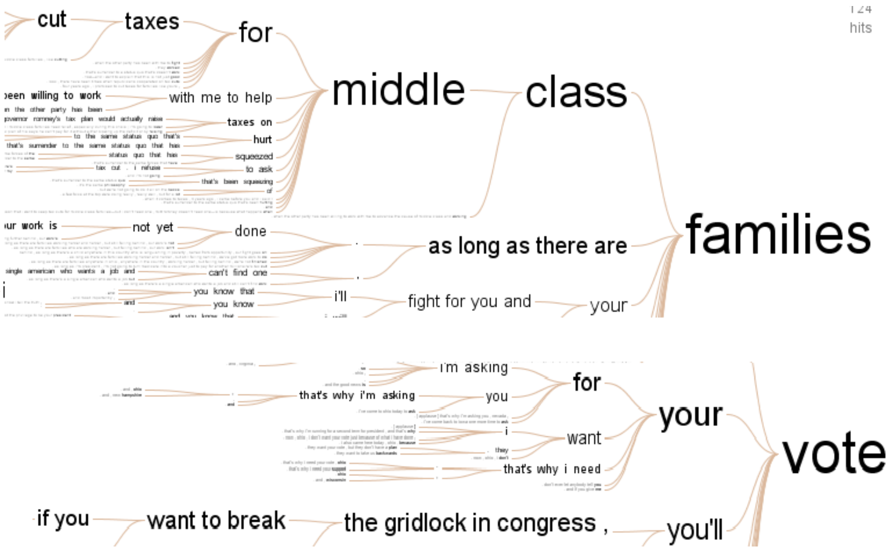


Figure 5: Word tree of Obama’s campaign Speeches focusing on tokens *families* and *vote*

Mitt Romney

An analysis of Mitt Romney’s word frequency using the same tool shows a distinctly different campaign strategy from the Obama one. Where Obama concentrated on the middle classes and focussed the debate on the domestic front, Romney’s campaign was of a defensive nature attacking and blaming the President the all the current ills of the country. He also made American foreign policy a key part of his campaign.

Romney’s word usage as seen in Figure 7 through to Figure 9 illustrates that he was appealing to the traditional Republican voter. He did not emphasise the united nature of the country but spoke about

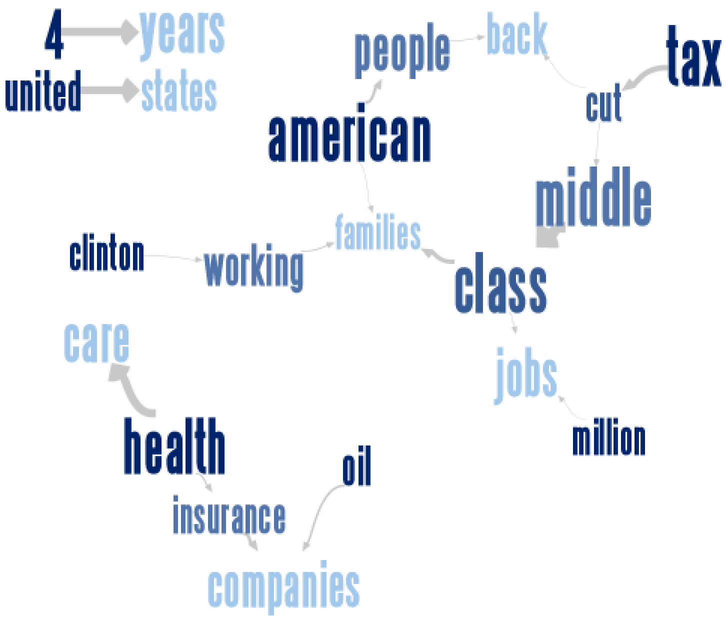


Figure 6: Phrase net of Obama's campaign Speeches

everything that was wrong with it. He lined up the enemies of the state, both domestic and foreign to appeal to his voters. Romney in his speeches concentrated solely on the President and his policies as key enemy No.1. In one case he said 'defeating him (Obama) is only one step forward.' He stated Obama's policies are not working, they are responsible for amassing 5 trillion dollars of debt, Obamacare is a bad policy, it affects jobs, Obama is a poster child for the arrogance of government. Romney hoped that his voters would agree that Obama was the main enemy on the home front while on the foreign front, Iran and Cuba were busy equipping themselves with nuclear weapons and that this proliferation of armaments stood to threaten the very existence of the United States.

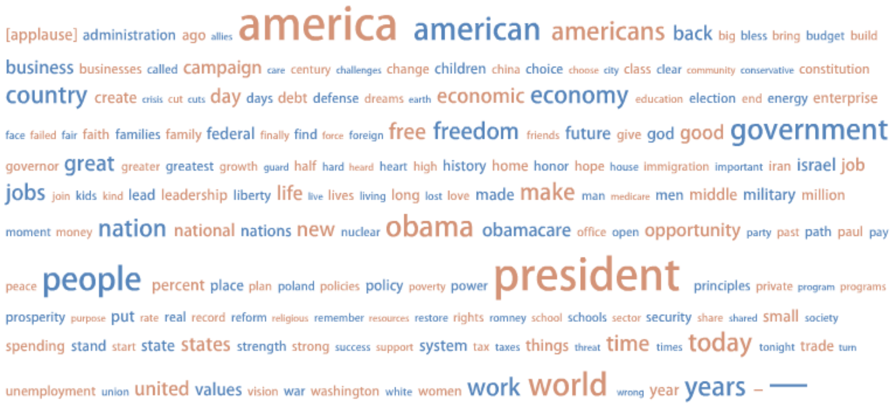


Figure 7: Tag cloud of Romney's campaign Speeches

5 Conclusion

This paper provided an analysis of some of the campaign speeches of the 2012 American Presidential Election outlining the fundamental differences between the Democratic and Republican parties. By utilising text mining applications such as text categorisation, similarity analysis and visual aids, the significant words and phrases used by Barack Obama and Mitt Romney were made apparent. The analysis

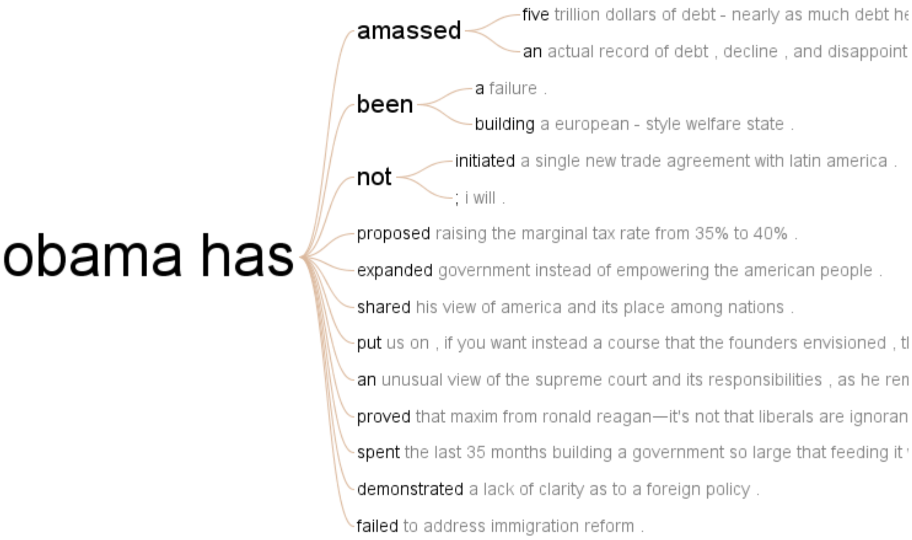

obama has
amassed
five trillion dollars of debt - nearly as much debt he
an actual record of debt , decline , and disappoint
been
a failure .
building a european - style welfare state .
not
initiated a single new trade agreement with latin america .
; i will .
proposed raising the marginal tax rate from 35% to 40% .
expanded government instead of empowering the american people .
shared his view of america and its place among nations .
put us on , if you want instead a course that the founders envisioned , tl
an unusual view of the supreme court and its responsibilities , as he ren
proved that maxim from ronald reagan—it's not that liberals are ignoran
spent the last 35 months building a government so large that feeding it
demonstrated a lack of clarity as to a foreign policy .
failed to address immigration reform .

Figure 8: Word tree of Romney's reference to President Obama

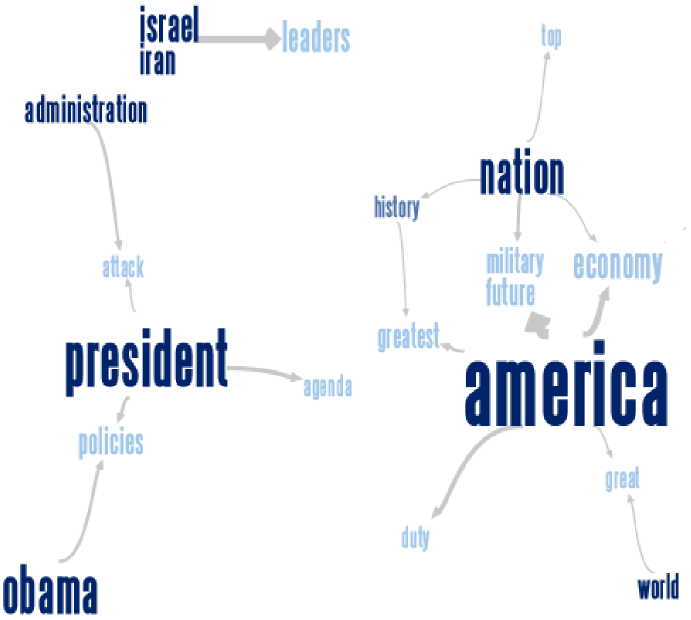


Figure 9: Phrase net of Romney's campaign Speeches

of the Obama/Romney speeches using the techniques above clearly identified the key policy issues for each side, the style of the respective campaign and the character of the strategy. In summation Obama focussed solely on the domestic economy, job creation and appealing to the middle classes, by stating ”*if you vote for me I will help you, I will fight for you in congress*”. He also emphasised the diverse nation of the country but that the strength of the country and the hopes of the future lay in unifying all peoples. His strategy was consistent and cohesive, focussing on the strength of a unified people in difficult economic times. The analysis presented here shows that he did mention many of the keywords identifying current problems in the country and mentioned them more frequently than his rival, demonstrating a perceived willingness to address these issues. Romney's strategy was to attack Obama and his policies from the outset. As a result of his target audience he jumped from issues of the threat of state assisted welfare states, to the nuclear threat posed by Iran and Cuba, to the need to build up a foreign defence system with allies such as Israel. He spoke about everything that was wrong with Obama and his policies and the deficit, but offered no concrete solutions. Romney was not proactive in his speeches but reactive to everything he perceived the Obama administration to be about. The dissimilarity in the speeches highlights the polar opposite views of both sides. It also illustrates the consistency of the Obama strategy and the fragmented and alienating nature of the Romney campaign, reinforcing the fundamental differences between both sides.

References

[1] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. Yale: Rapid prototyping for complex data mining tasks. In Lyle Ungar, Mark Craven, Dimitrios Gunopulos, and Tina Eliassi-Rad, editors, *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 935–940, New York, NY, USA, August 2006. ACM.

[2] Murphy Choy, Michelle LF Cheong, Ma Nang Laik, and Koo Ping Shung. A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *arXiv preprint arXiv:1108.5520*, 2011.

[3] Scott P Robertson. Changes in referents and emotions over time in election-related social networking dialog. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pages 1–9. IEEE, 2011.

[4] Adam Bermingham and Alan F Smeaton. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1833–1836. ACM, 2010.

[5] Kevin Coe. George w. bush, television news, and rationales for the iraq war. *Journal of Broadcasting & Electronic Media*, 55(3):307–324, 2011.

[6] Marco R Steenbergen, André Bächtiger, Markus Spörndli, and Jürg Steiner. Measuring political deliberation: a discourse quality index. *Comparative European Politics*, 1(1):21–48, 2003.

[7] Robert Klemmensen, Sara Binzer Hobolt, and Martin Ejnar Hansen. Estimating policy positions using political texts: An evaluation of the wordscores approach. *Electoral Studies*, 26(4):746–755, 2007.

Predicting Premiership Football Match Results

Machine vs. Men – Can predictive models outperform human experts?

Elaine Kirwan[1], Markus Hofmann [2]

[1] Institute of Technology Blanchardstown, Dublin 15, Ireland
leabharnua@gmail.com

[2] Institute of Technology Blanchardstown, Dublin 15, Ireland
markus.hofmann@itb.ie

Abstract

Predicting the outcome of football matches is a popular research area primarily due to the popularity of the sport worldwide and it is therefore unsurprising that there have been numerous previous attempts both commercially and academically to forecast Premier League match results with varying degrees of success. In this research paper we apply the CRISP-DM (CRoss-Industry Standard Process for Data Mining) methodology to create and explore a variety of predictive classification models in an attempt to accurately and objectively predict the outcome of English Premier League football matches using freely available online data that is based only on previous Premier League match results from 2002 to 2012. Using an iterative approach, numerous data visualisation, pre-processing and transformation techniques have been applied and evaluated to determine the most effective predictive model, which has been assessed by its ability to correctly forecast the match outcome of the final 50 matches in the 2011/2012 Premier League season as per the initial business and data mining objectives. The result of this study showed that a number of models outperform human expert match predictions.

Keywords: Soccer predictions, predictive modelling match results

1 Introduction

Football is widely acknowledged to be the most popular sport in the world and as a result, in recent years vast amounts of football data with varying levels of detail has for the first time been collected and made publicly available online in various forms such as match results, in-play events, individual player and team performance statistics and betting odds. In the past, football enthusiasts and analysts were entirely reliant on their own team knowledge, expertise and experience to make what was essentially an educated guess at football match, league and championship outcomes. However by combining the global popularity of football with the now widely available historical match data, the football prediction landscape has significantly changed as the potentially biased instinct based approach has been replaced by a more objective analytic process [1, 2]. It is therefore unsurprising that Premiership match and league forecasting has recently attracted a lot of attention and has also experienced a huge growth with the introduction of numerous fantasy leagues which are purely for entertainment to the more serious enthusiasts using various online forecasting algorithms, applications and advanced predictive analytics in an attempt to achieve an edge over the bookmakers. Some mainstream online examples of Premier League football prediction applications would be the Castrol Predictor [3], pi-football.com [4] and SoccerWinners.com [5].

This paper aims to investigate through the application of the CRISP-DM methodology and using detailed historical premiership match information from 2002 to 2012, what the most important statistical features of the data are and if a match outcome can be accurately predicted using data mining. As a football fan, the opportunity to potentially forecast the outcome of football matches which are traditionally considered to be inherently unpredictable is equally appealing and challenging. Therefore the purpose of this exploration is to demonstrate the use of predictive analytics by developing a classification model that is capable of accurately predicting the results of Premier League fixtures into one of three possible outcomes – home win, away win or draw. A secondary goal is to improve upon the prediction accuracy of one of today's most popular football pundits Mark Lawrenson, who publishes weekly Premier League match predictions online on BBC Sport [6].

2 Objective

The objective of this study is to accurately predict the outcome of Premier League football matches using a freely available online data set that is based on past English Premier League match results from the 2002/2003 to 2011/2012 seasons. The classification model should identify and consider the relevant in-play information such as shots on goal, number of home or away goals scored, the impact of yellow or red cards received along with each team's recent home and away form in order to forecast the match outcome into a pre-defined number of possible results. The goal is therefore to introduce consistency and objectivity into the match forecasting process for football fans or betting enthusiasts so that they can more effectively predict match outcomes to within a reasonable degree of accuracy, which we will measure by comparing our classification results with the predictions of the acclaimed football pundit Mark Lawrenson on BBC Sport online [6].

Football match result prediction presents an interesting research problem as there are numerous important factors that can impact upon the outcome of a competitive match such as player availability (through injury and suspension), team selection, additional European games, coaches' in-play tactics and match officials performance and decision making. These dependant considerations all combine to make accurate result forecasting difficult when using only previous results and scores for data mining.

Although there is a significant amount of Premier League match data available, meetings between certain teams are infrequent and in the case of some recently promoted teams non-existent, therefore limiting the data mining models' ability to generalise the resulting classification for future match prediction. Additionally the team structure can vary from season to season through player sales and acquisitions and also through management changes which can make previous result information irrelevant within a short period of time and this is why we use only 10 years of historical data. Later in this paper, we will experiment with applying greater weighting and ranking to more recent information in an attempt to deal with this problem.

3 The Data

The Premier League result data used in this research paper is provided by www.football-data.co.uk [7] which is an independent website that collects and presents data from over 20 European leagues as far back as the 1993/1994 season. The match data from 2002 until 2012 have been combined into a single comma separated value file containing approximately four thousand records and it is made up of half time and full time scores and results and also various match statistics including the number of home or away goals scored, shots on goal, fouls, bookings and referee names. As previously discussed, due to the highly unpredictable nature of football matches, the data understanding exploration phase is especially important to identify the key attributes and features that can be used to detect any interesting patterns and trends within the data set Table 1 shows the available attributes.

4 Data Exploration

Football is traditionally a low scoring sport with an average of just 2.6 goals per match and this combined with the low average numerical values for key events during a match, makes identifying attributes or events that have a direct influence on the final outcome extremely difficult.

4.1 General Exploration

The first pattern of note relates to the distribution of the full time result in favour of matches resulting in a home win with about 47% of all matches ending in a win for the home team, demonstrating a significant home advantage. The remaining results are almost evenly divided between draws and away wins, at 26% and 27% respectively.

Table 1 Data Attributes

Attribute	Description
Div	Premier League Season
Date	Match Date (dd/mm/yy)
HomeTeam	Home Team
AwayTeam	Away Team
Referee	Match Referee
Age	Match Date
HTR	Half Time Result (H=Home Win, D=Draw, A=Away Win)
FTR (Label)	Full Time Result (H=Home Win, D=Draw, A=Away Win)
FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HC	Home Team Corners
AC	Away Team Corners
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards

Research suggests that home advantage can be attributed to a number of factors such as local crowd support, pitch/stadium familiarity, referee bias favouring home team decisions, away team travel fatigue and increased home team player motivation [8] and in Figure 1 using time series analysis we can visually assess the effect of home advantage based on total points earned in the Premier League, which has remained consistent over the past ten years.

Although there is a clearly uneven distribution of outcomes, for data mining purposes there should be enough examples of each outcome (label) in the football data set, however we will experiment with bootstrap sampling techniques to see if any predictive accuracy improvements can be achieved.

Although there is a clearly uneven distribution of outcomes, for data mining purposes there should be enough examples of each outcome (label) in the football data set, however we have experimented with bootstrap sampling techniques to see if any predictive accuracy improvements can be achieved.

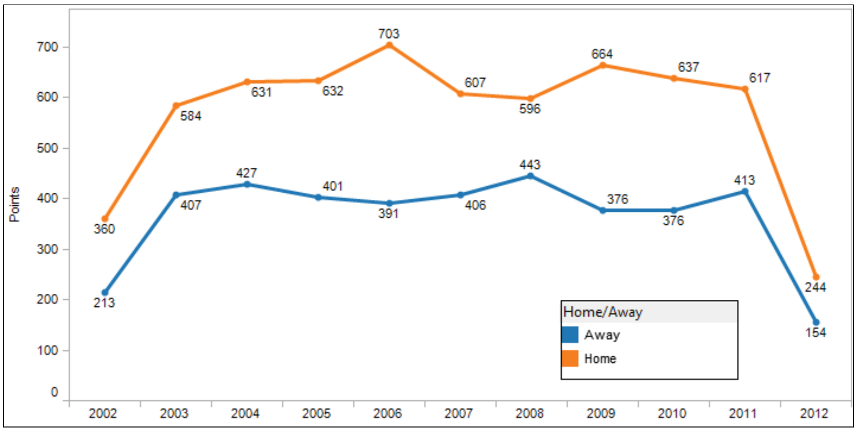


Figure 1. Total points won home and away (by year)

4.2 Recent Form Analysis

A common analytical consideration when predicting match outcomes is to investigate the recent form of the two teams involved both at home and away (Figure 2). Recent form is normally measured in footballing terms by looking at the results of the last six matches but we are actually interested in analysing any long term (full season) and short term (week on week) performance trends so form could be defined in a number of other ways:

- Form based on more or less than six matches
- Form based on the entire season (or more)
- Weighted Form through exponential decay, with most recent matches receiving the highest weighting (more on this later in the paper)
- Home and Away Form

Each of these definitions was considered and applied during the data pre-processing and modelling phases to identify the most suitable method to achieve our data mining objectives.

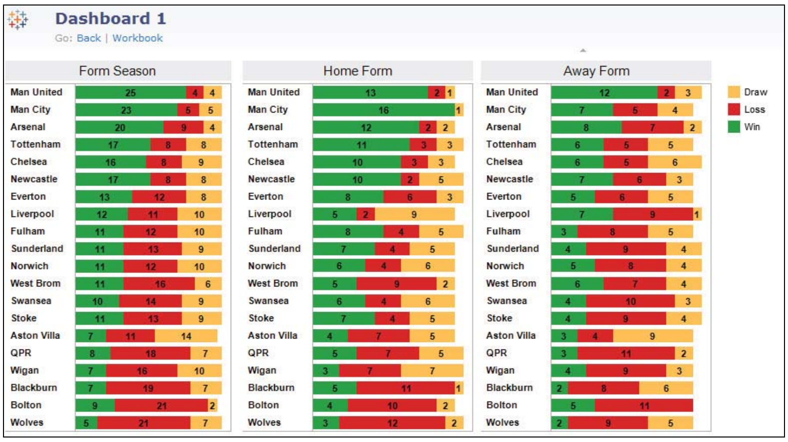


Figure 2. Recent Form Guide–Stacked Bar Chart Season Form

4.3 Goal Difference

By creating a current League Table we can see that there is a clear linear relationship between goal difference and league position with the league leading teams displaying a superior goal difference (see Figure 3). This is further proven by examining the resulting correlation coefficient which at 0.981 implies that a team's record of goal scoring (attack strength) and conceding (defensive strength) is a very good indicator of a team's likelihood and ability to win the match. The effect of incorporating goal difference and point values into our training and test data sets will be analysed in more detail during the pre-processing and predictive modelling phases.

4.4 Number of Yellow/Red Cards issued

The number of red cards issued unsurprisingly has an impact on the overall outcome of a match, but what is interesting is the negative effect is substantially more pronounced on the home team, with a clear reduction in home advantage from 47% to just 23% likelihood of winning (also see Figure 4):

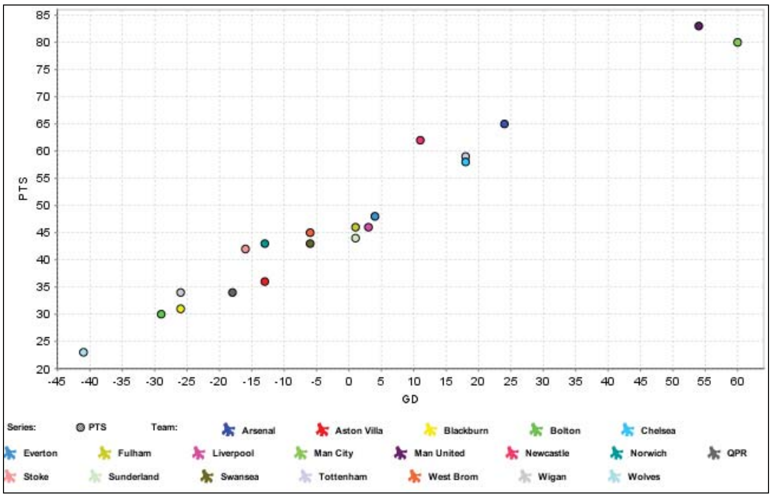


Figure 3. Total Points vs. Goal Difference 2011/2012 (stopped at 24/04/2012)

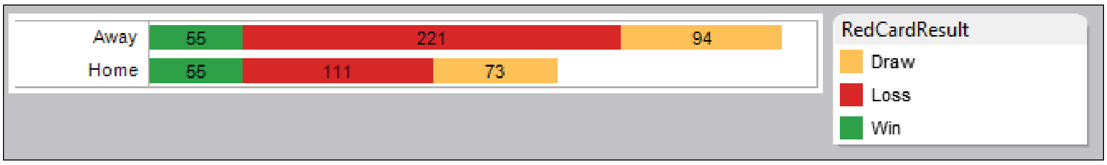


Figure 4. Impact of Red Card on Match Outcome (Home/Away)

Using a kernel density plot in R we can compare the effect of yellow cards on the home and away team grouped by match outcome. We note that on average more yellow cards are issued to the away team (1.8 versus 1.3 to the home team) and also by looking at the distribution shape of match outcome we can see that there is a higher density of away wins when 1 or more yellow cards have been issued to the home team with a more even effect when yellow cards are issued to the away team.

5 Data Processing

5.1 New Attribute Generation

A number of new attributes were introduced and incorporated into the source Premier League data set and each will be assessed individually and in combination with others to determine the best possible classification model. Each one has been identified as having possible significant predictive capabilities through our earlier data visualisation processes and they have been created through a series of pre-processing techniques

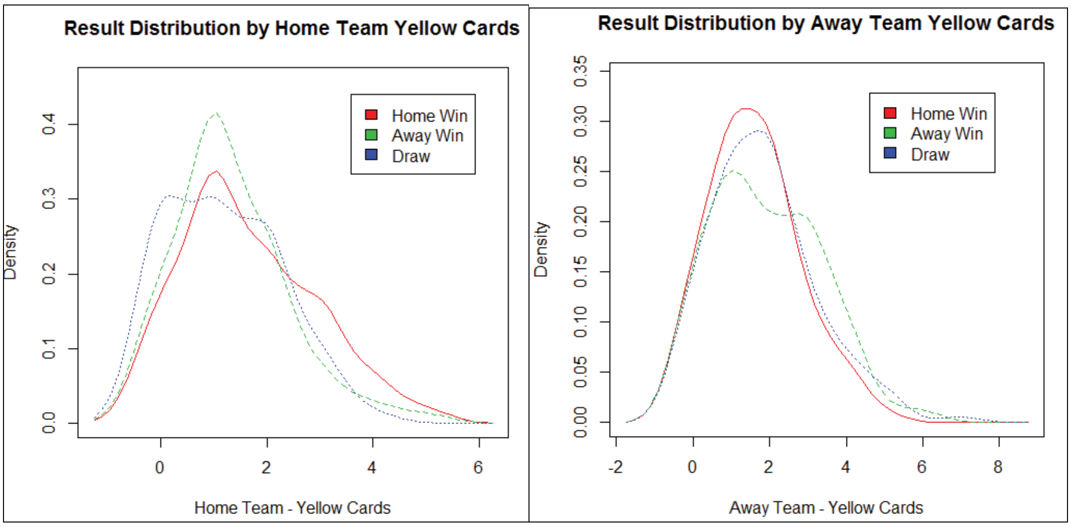


Figure 5. Kernel Density Plots – Impact of yellow card on Distribution of Outcome

5.2 Age

In order to create Match “ageing” for recent form analysis and weighting, we have created a new Age attribute by calculating the difference in days between the match date and today’s date.

As identified previously, current form can have an impact on match outcome so the Age attribute is required to vary the time periods that points and goal difference are included for, such as the previous six matches only, full season etc. Age was used to filter time periods only and excluded from the data mining analysis itself as it is fully correlated with date.

5.3 Total and Home/Away Points and Goal Difference

Through a series of data transformations and pre-processing, we split the match results into various important Home and Away form statistics. Using a number of aggregations, pivots and value mappings the league points and goal differences were calculated. We have created both Total and also Home/Away points and goal difference values in order to evaluate their predictive effectiveness later in the modelling phase. The full list of new attributes establishing team form is outlined in Table 2.

5.4 Home/Away Performance Rating:

As a result of earlier analysis in relation to influence of form, we introduced a new home and away rating attribute that calculates the likelihood of a home win, away win or draw by using the following simple formulae:

$$HWR = (HW+AL)/(HW+HD+HL+AW+AD+AL)$$
$$AWR = (AW+HL)/(HW+HD+HL+AW+AD+AL)$$

DR = (HD+AD)/(HW+HD+HL+AW+AD+AL)

These attributes were calculated and added to the dataset as shown in Table 2.

5.5 Exponential Decay

One of the most important features of this data mining analysis is the application of increased weighting for more recent fixtures as we have seen how form can contribute greatly to match outcome. Therefore exponential decay used in this context is the valuation of recent matches by heavier weights than older matches with a decreasing weight (decay) the older the game. This ageing can be achieved in a number of ways and we began with weighting the most recent season's results using Bootstrap Sampling where Season 2011/2012 receives full weighting (Sample Ratio = 5), 2010/2011 (Sample Ratio = 4) and 2002 – 2006 are not bootstrapped at all.

Table 2: Newly Generated Data Attributes

Attribute	Description
HPTS	Home Points
HGD	Home Goal Difference
HGF	Home Goals For
HGA	Home Goals Against
HPLD	Home Matches Played
HW%	% of Home Match Wins
HD%	% of Home Matches Drawn
HL%	% of Home Match Losses
HTPTS	Home Team Total Points (Home and Away)
HTPLD	Home Team Total Matches Played (Home and Away)
HTGD	Home Team Total Goat Difference (Home and Away)
APTS	Away Points
AGD	Away Goal Difference
AGP	Away Goals For
AGA	Away Goals Against
APLD	Away Matches Played
AWX	% of Away Match Wins
AD%	% of Away Matches Drawn
AL%	% of Away Match Losses
TPT5	Away Team Total Points (Home and Away)
ATPLD	Away Team Total Matches Played (Home and Away)
ATGD	Away Team Total Goal Difference (Home and Away)
HWR	Home Win Performance Rating (%)
DR	Draw Performance Rating (%)
AWR	Away Win Performance Rating (%)

Another possible approach was to only consider the most recent records removing older results completely from the analysis and we can vary this using the Div (Season) and Age attributes. The effect of this form weighting process on predictive accuracy is analysed in more detail during the following Modelling section.

6. Modelling

In this paper we selected two classification models for evaluation (1) k-NN and (2) Naïve Bayes. We began by looking at the baseline performance of both Naïve Bayes and k-NN.

Parameter optimisation processes identified k=17 to be the optimal value using 10-fold cross validation.

In the context of the football match classification model, there are six possible outcomes as per the cross validation confusion matrix:

- True Home: Home win is correctly classified as a win for the home team
- False Home: Away win or draw is incorrectly classified as home win
- True Away: Away win is correctly classified as a win for the away team
- False Away: Home win or Draw is incorrectly classified as away win
- True Draw: Draw is correctly classified
- False Draw: Home win or away win is incorrectly classified as draw

As the confusion matrix provides total predictive accuracy where false positives and false negatives are assumed to have equal value we also evaluated our models in terms of a cost matrix, where each outcome has an associated cost to ensure that costly misclassifications are minimised. As a result we evaluated the overall classification performance based on a combination of total accuracy and cost.

In order to calculate cost, we assumed that our model is being used to assist in a simple betting decision making process where each correct bet is worth €10 in winnings but the cost of placing a home win bet is €5 (2-1 odds) and the cost of placing an away win and draw bet is only €2 (5-1 odds) as the chances of away win and draw outcomes are significantly lower. Therefore a correct home win bet is said to be worth + €5 (€10 winnings - €5 cost of bet) and a correct away win and draw result is worth €8. An incorrect home win bet results in a loss of - €5 but an incorrect away win and draw bet would only result in a loss of -€2.

6.1 KNN

The standard k-NN classification model in RapidMiner [9] with no data pre-processing has provided mixed results. While its ability to correctly classify the occurrence of a home win at 75.25% is relatively high, the overall accuracy and cost per bet at 45.13% (+/- 2.07) and €0.39 respectively are quite disappointing but unsurprising.

Some model refinement process dealt with the outlier values identified in our earlier analysis by binning values using discretisation on a subset of the numerical attributes (Home/Away Shots, Home/Away Shots on Target, Home/Away Corners and Home/Away fouls). At each iteration of k-NN, we also re-assessed the optimal value for k to ensure the best possible outcome.

While the resulting confusion matrix shows a considerable improvement in all classifications in terms of both recall and precision, the total accuracy is still relatively low at 61.86% due mainly to the difficulty in correctly identifying the draw outcome. As there are considerable fewer instances of draw and away win outcomes in the data set and because the data set is relatively small containing only 4,000 records in total, Bootstrap sampling was applied with a sample ratio of 5 to simulate a larger data set that is representative of the original data.

The standard bootstrap operation yielded sizable improvement in both accuracy and cost and inspired confidence in an even better classification result following the application of the exponential decay model described earlier in the paper. We attempted a number of different configurations including varying the seasons and bootstrap ratios. The best model identified using the exponential decay design included bootstrap weighting the 2011/2012 – 2007/2008 seasons and to exclude any seasons prior to 2007 from the bootstrap sample but the best result achieved with this form weighting model and k-NN actually resulted in a lower accuracy (67.79%) and cost return (€3.15) than the standard bootstrap sample which was disappointing.

The next step was to assess the effect of incorporating the Total and Split Home/Away form into the model and also to identify the best possible configuration in terms of how much form should be included (i.e. last 6 matches, full season). This did not yield any improvements.

6.2 Naïve Bayes:

Taking a similar approach for the Bayes classification model as we did with k-NN, we began by dealing with the outlier values. As the normalization and discretization cross validation analysis carried during early stages of this project yielded similar results for discretization by entropy and proportion transformation, the effects of both methods were assessed. The original data set that has again been bootstrapped with a sample ratio of 5.

The results are almost identical but in most cases the Discretize by Entropy algorithm performs marginally better when predicting home and away wins (Accuracy of 83.3% +/- 0.96) and therefore discretization is selected as the most appropriate method to deal with outlier values when using the Naïve Bayes classification model.

Again we assessed the Naïve Bayes model by using the same three advanced models as with k-NN and by applying similar data transformation processes using the Total Points and Goal Difference, Home and Away Points and Goal Difference and the Home and Away Performance rating models using cross validation to identify which returns the best accuracy and cost.

The optimal configuration for Bayes is also using the split home and away points and goal difference classification model and it achieved a total predictive accuracy of 83.82%

7 Evaluation & Conclusion

The modelling phase was originally undertaken with only classification accuracy and cost in mind, but the biggest mistake made was that consideration should have been given earlier to running the unknown test examples through the model to determine actual predictive accuracy with the unknown numerical values. Initially very encouraging results were in reality found to be nowhere nearly as good when evaluated with real test match fixtures. In addition the assumption that the optimize parameter operator always yielded the best value for k in k-NN model assessment was a little misguided as again this value was founded in a cross validation assessment that presumed that all numerical values were present within the test data set. In terms of predictive accuracy within the confusion matrix the k value was correct but when applied to our 50 test Premier League matches, it was found that a higher value for k actually worked better. Due to project time constraints, we subsequently evaluated k values in increments of 5 and found that k = 20 performed the best with the Home and Away Points and Goal Difference model giving a maximum actual predictive accuracy of 53%. It was also interesting to see how both models performed on the real match day test data sets as through cross validation k-NN indicated a lower predictive accuracy than Naïve Bayes (67.79% vs. 83.46%). However in practice they actually performed almost identically with k-NN successfully predicting only one additional match correctly (27 vs. 26 matches correctly predicted).

Throughout the CRISP-DM process it became clearly evident that there are a huge number of variables and factors that must be considered when attempting to predict Premier League match outcomes and to achieve the maximum possible predictive accuracy through data mining there are perhaps a number of additional attributes that should be included in the analysis to truly reflect the intricacies of a football match such as goals times, goal scorers, indication of penalty and own goals scored, match attendances and team line-ups for example to ensure as comprehensive a study as possible. For the purpose of this project, we chose to focus solely on previous match outcomes for the sake of free data availability and model simplicity but even still, we have proven that the Premier League matches do have a certain predictable element.

It is also worth noting that there has been a small element of bad timing as the 2011/2012 season is widely considered to have been the most unpredictable since the Premier League was first established 20 years ago. In the last 8 weeks of the season especially there have been a number of "shock" results as the bottom three teams have struggled with relegation and with an equally unanticipated battle at the top of the league for Champions League qualification next season. The last 50 matches of the 2011/2012 season were selected for model evaluation and in hindsight that may not have been wisest decision as the dramatic climax to the season have meant that these particular matches have arguably been amongst the most unpredictable matches ever. For example, there were a number of matches where neither the human expert nor the classification models were able to correctly predict the match outcome of what are considered to be surprising results - Wigan beating Arsenal at home on the 16/04/2012 or West Brom beating Liverpool at home on 22/04/2012.

Even so, as a result of this study, we have produced two predictive classification models that both outperform a human expert when forecasting Premier League match outcomes where k-NN and Naïve Bayes achieved 53% and 51% respectively with Mark Lawrenson correctly predicting only 47%. Perhaps more importantly however, both algorithms are capable of objectively and without bias forecasting the match outcome and for this reason we can conclude that the original objective has been successfully achieved.

References

[1] R. Schumaker, O. Solieman, and H. Chen, Sports Data Mining, Springer, 2010.

[2] R. Schumaker, O. Solieman, and H. Chen, "Sports Knowledge Management and Data Mining," Annual Review of Information Science and Technology (ARIST), Volume 44, 2009.

[3] Castrol Grand Prix Predictor. [online] Accessed: 29 November 2012. Available: http://predictor.autosport.com/ .

[4] Pi-football. [online] Accessed: 29 November 2012. Available: http://pi-football.com/ .

[5] Soccer Winners. [online] Accessed: 29 November 2012. Available: http://www.soccerwinners.com/ .

[6] BBC Sport, 2012. Mark Lawrenson's Premier League predictions. [online] Accessed: 29 November 2012. Available at: http://www.bbc.co.uk/sport/0/football/17852284

[7] Historical Football Results and Betting Odds Data, 2012. Premier League FT & HT results; match stats; match, total goals & AH odds [online] Accessed: 29 November 2012. Available at: http://football-data.co.uk/englandm.php .

[8] Koyama, M., Reade, J., 2008. Playing Like the Home Team: An Economic Investigation into Home Advantage in Football, International Journal of Sport Finance, 4 (1), p. 16-41.

[9] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler: YALE: Rapid Prototyping for Complex Data Mining Tasks, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06), 2006

Session 4

Web & Cloud Technologies

A Performance Analysis of WS-* (SOAP) & RESTful Web Services for implementing Service and Resource Orientated Architectures

Philip Markey, Gary Clynch

Department of Computing, Institute of Technology Tallaght, Dublin 24
phil.markey@gmail.com
gary.clynch@ittdublin.ie

Abstract

The past number of years have seen the emergence of Service-Oriented Architecures as a dominant architecture for implementing enterprise scale distributed systems. Two main styles of SOA exist, namely SOAP based services and RESTful services. There has been much comment and debate on the pros and cons of each approach to implementing a SOA, a lot of which has surrounded the performance characterictcs of both approaches.

In this paper, the authors presents the results of a performance analysis that was conducted on a set of test SOA scenarios implemented using both SOAP and RESTful approaches; in particular the caching capabilities of REST have been exploited with significant benefits accruing, an option not available with SOAP based approaches.

Keywords: SOAP, REST, RESTful, Service Oriented Architecture, Resource Oriented Architecture.

1 Introduction

In distributed computing Service Oriented Architectures (SOA) and Resource Orientated Architectures (ROA) have been viewed as competing architectures for implementing an enterprise scale distributed system using web services [10]. The first architecture, SOA, is associated with SOAP based web services and the WS-* stack of specifications. The second architecure, ROA, is associated with the Representational State Transfer (REST) based web services.

Both approaches are alternative architectures for implementing a distributed system with pros and cons associated with each. For example it has been suggested that the complexities of the WS-* stack could be viewed negatively when looking towards the comparatively simplicity of the RESTful approach. The opposite could also be surmised as, the simplistic use of the REST, may not be sophisticated enough to be used with an enterprise system and therefore the mature and well documented WS-* stack should be utilised in its place. It would be interesting in particular to compare the performance of both approaches.

This paper documents the results of a performance analysis performed on a set of benchmark web services implemented using both SOAP-based and RESTful approaches.

2 SOA/ROA & Web Services

SOA is a flexible set of design principles, when utilised efficiently can provide a loosely coupled set of services via a single endpoint, that can form one, all-encompassing application that simplifies machine-to-machine communication [6]. For this SOA services utilise a set of standards that allow applications to be published, discovered and invoked such as the Web Services Description Language (WSDL) [3].

SOA services focus is on providing a schema and message-based interaction with an application [3]. XML is the common message format that is used for SOA services; the SOAP standard is normally adopted containing an XML payload [6].

At a fundamental level, the SOAP-based architecture revolves around the transmission of XML-encoded messages over a transport medium and is utilised in the formal Web Service approach. The specifics of a SOAP service is that it follows a very well defined set of rules that are prescribed for in the WSDL files, which are essentially XML structured files adhering to a W3C-specified grammar [5].

ROAs concepts are based on three points of Resources, URI and Representations, which are provided over multiple endpoints, each linking to a single resource [7]. Resources are objects that can be referenced and stored on a computer; this could be the returning result from a SQL query or a bit stream representation of a document. Representations are simply the state/format of data of the required resource.

The Resource is referenced by the use of a URI [1]. The URI is essentially just the name and address of the requested resource, with best practices should be as descriptive as possible to best match the resource that is being addressed.

REST is a set of specific guidelines that can be covered to produce an implementation of a RESTful service [7]. It can be used to design a Web Service that focuses on system resources, including how resource states are addressed and transferred over HTTP [8].

REST is based on a set of transfer operations that are universal to any data storage and retrieval system. These operations are commonly referred to by the acronym CRUD as shown in *Table 1: Core HTTP CRUD Methods*, which stands for Create, Read, Update, and Delete.

CRUD	**REST**	
CREATE	POST/PUT	Initialise the state of a resource at a given URI
READ	GET	Retrieve the current state of the resource
UPDATE	PUT	Modify the state of the resource
DELETE	DELETE	Clear a resource.

Table 1: Core HTTP CRUD Methods

3 Benchmark – Experiment

A set of benchmark experiments were conducted in order to compare and contrast the performance of the SOA and ROA alternatives to implementing a distributed system using web services. The web services were implemented using WCF and the Web API for the .Net platform in C#. The performance metric measured was network weight i.e. the amount of network traffic that resulted from an interaction between a client and the web service.

3.1 The Benchmark Environment

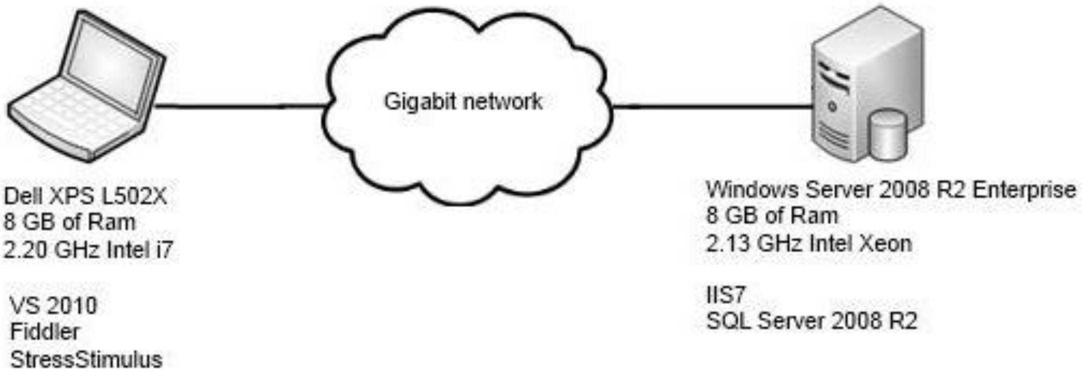


Figure 1: Benchmark Network Topology

As illustrated in Figure 1, the benchmark environment is a simple topolgy, that consists of a Gigabit network, a server with a Windows Server 2008 R2 Enterprise installation and a Dell XPS. Within this network set up, the server acts as the Web Service host, hosted in Internet Information Services 7 (IIS7). The server is also running an instance of SQL Server 2008 R2, which is accessible to the Web Services.

On the XPS there is an intermediary piece of software installed that acts as a proxy host, called Fiddler2 [4]. With Fiddler2, HTTP traffic passes through it for traffic in both directions and has the capability to view this traffic as it passes through and provides the ability to be able to inspect the traffic.

To further extend the capabilities of Fiddler2, a plugin called StresStimulus [9] is installed; StresStimulus is a load test and performance tool. Using StresStimulus it is possible to record a number of traffic scenarios and then replay the same scenarios under various load patterns, with a varying number of virtual users while monitoring how the site is performing under that load.

In the benchmarking scenarios used in this session, Fiddler2 is used to capture and analyse the traffic and then StresStimulus is used to replay the same request traffic over again for each repeated test. In all test scenarios only one virtual user is ever used.

3.2 Benchmark Scenarios

For the benchmarking session there are two scenarios that are carried out to allow a comparison to be drawn.

In the first scenario, the web service calls are used to request an object from each of the respective Web Services. The object that is requested is generated from a record held on the database, which is located on the test server. The item is a simple country record, that is made up of 11 fields, that, when populated represent a country. These fields include items such as the country acronym, which would be a three character string in the manner of ‘ire’, then the name of the country ‘Ireland’ and the database primary key, which is an integer, such as 103 in this case, amongst others. The fields of the country record are primarily basic types of String, Boolean and DateTime and are requested within the Web Services by the use of the Entity Framework. This scenario is used to demonstrate the size and weight of each request response and demonstrate the affect the wrapping protocol can have on each of them.

The second scenario comprises of two parts, part one takes a similar format as the first scenario (as described above), wherein each service call will be requesting an object from its respective service, the service will be fulfilling this by use of a call the database. For this scenario the request is for a collection of country objects as previously described, the collection will contain 249 items in total.

Instead of solely making one call, this test call will be repeated over a set timeframe of one minute per test, with each call being placed consecutively. This would demonstrate not just the weight of the call and its associated wrapping protocol, as was demonstrated in the previous scenario but also the amount of traffic that can be produced by each service, when an iterative call is placed for a large response. This can then also show how many calls can be successfully placed with each of the services. This will then be further extended for the second part to this scenario, which will enable caching; as can be utilised when using a RESTful approach to the service call.

In the second part of this scenario, which is only utilisable by the RESTful services, as a RESTful service call uses the Web API and the HTTP Headers fully and in particular the GET Header. The responses from the service can then be cached, unlike the SOAP counterpart that can only uses an over loaded POST Header; the POST header is deemed to be unsafe and therefore non-cacheable, as changes, such as update or deletes, can be sent to the service and therefore to the database with the request.

In this secondary section, the test is run again, calling for the same collection of 249 country objects over a one minute timeframe, this time with the caching abilities enabled. From this, it will again show the weight of the traffic that is transmitted and show the number of calls that could be successfully placed within the given timeframe. Further to this, to show what affect enabling caching can have when looking at services request/responses.

Each of the separate test scenarios were executed four times and from this the average returned results were used for the comparison. The numerous executions were used to reduce the possibility of an erroneous call or network traffic swaying the results in either direction, for any of the tests, therefore having a genuine reflection of the result for each individual scenario.

4 Results & Analysis

4.1 Scenario One

In the first scenario, both of services had an operation invoked that interacted with a database and retrieved a single country object. This country object was then packaged up and transported back to the client in the various formats; SOAP for the WS* style, and XML and JSON for the RESTful style. This was used to demonstrate the differences in size of the sending and returning single payloads.

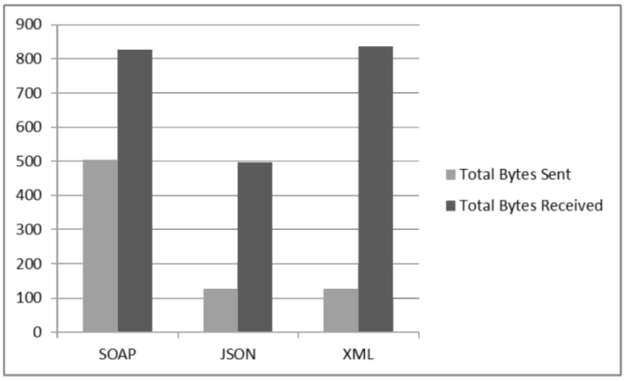


Figure 2: SOAP vs. REST Single DB Item Comparison

	SOAP	JSON	XML
Total Bytes Sent	505	128	127
Total Bytes Received	826	496	836
Total Overall	1331	624	963

Table 2: SOAP vs. REST Single DB Item Comparison

Within this test, shown in Figure 2 and Table 2 differences can be seen as to the comparative between the SOAP and RESTful calls. SOAP sending 505 bytes, whereas both the JSON and the XML RESTful calls only sending 128 bytes and 127 bytes respectively, this representing only 25% the size of the SOAP request. On the received data, the XML RESTful call returns an extra 10 bytes compared to the SOAP response but the JSON call received little over half this with only 496 bytes.

In the overall data transported for this test case, the data sent/received by the RESTful XML service is 72% that of the size of the SOAP service but the JSON service is smaller again, with a total data transported representing just 47% that of the SOAP service and 65% that of the XML service.

4.2 Scenario Two

Scenario two comprised of two parts, firstly both types of service were used to call a method that would return a collection of items from the database (249 items per call) repeatedly over a minute duration, with each call being placed consecutively. After this the test was carried out a second time for the RESTful service, this time with caching enabled.

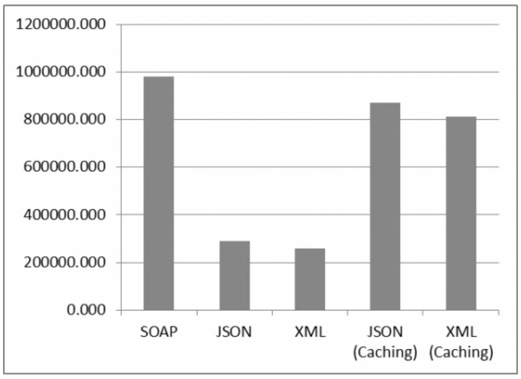


Figure 3: SOAP vs. REST DB Collection Comparison - Bytes Sent/Sec

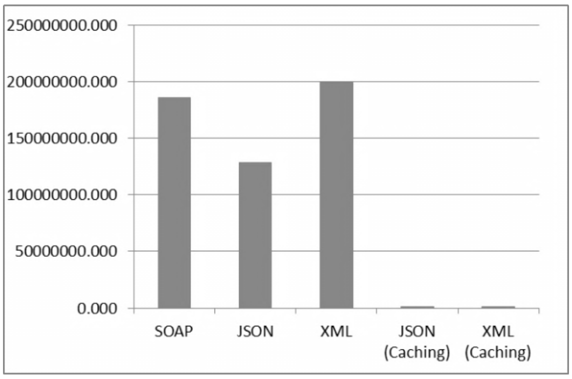


Figure 4: SOAP vs. REST DB Collection Comparison - Bytes Received/Sec

	SOAP	JSON	XML	JSON (Caching)	XML (Caching)
Total Bytes Sent	981682.5	289230	260022	868937	810619
Total Bytes Received	185677951.25	128586060	199181080	961354	1054510
Total Overall	186659633.75	128875290	199441102	1830291	1865129

Table 3: SOAP vs. REST DB *Collection Comparison*

From Figure 3, Figure 4 and Table 3 we can see that the weight of the calls from the SOAP service is relatively heavy in comparison to the RESTful calls and the responses from the XML RESTful service is yet again marginally heavier than it SOAP counterpart as previously noted.

When we take a view of the total bytes sent/received for each call service, it can be noted that in the scenario where caching is not enabled, the RESTful XML call is in fact larger when compared to the SOAP service call by 6.8%, with the JSON service still the smallest at 64.6% the total size of the XML call and this representing 69% that of the SOAP service.

The interesting difference comes with caching enabled. Although the RESTful requests increase, they are still significantly less than that which is required by the SOAP service. With this, the responses are reduced by a distinct margin due to the server only having to return a header message, which states that no change has occur to the data and a suppressed body. This in turn changes the total data transported to show that with caching enabled the RESTFul calls are 0.9% that of their SOAP counterpart.

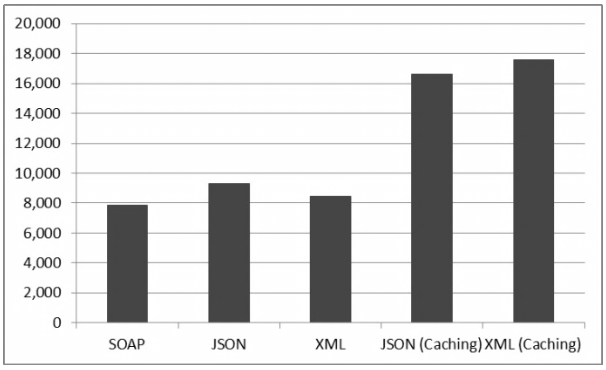


Figure 5: SOAP vs. REST Total Requests Send/Received

	SOAP	JSON	XML	JSON (Caching)	XML (Caching)
Total requests	7,885	9,330	8,456	16,632	17,559

Table 4: SOAP vs. REST Total Requests Send/Received

Also, if we are to look at the total amount of requests that are being sent and received during the minute test, as shown Figure 5 and Table 4, both the services were able to process a similar number of calls without caching enable. When the caching was enabled for the RESTful service calls, the service is able to process approximately double the amount from the previous iteration.

4.3 Analysis of Results

With both SOAP and REST termed under the same heading of Web Services, they can both stand separately with their architectural styles for their uses in SOA and ROA respectively and each can have significantly differing effects on a network load.

A major point for utilising a RESTful service is the fact that a RESTful service follows the Web API and uses the appropriate headers of GET, POST, PUT and DELETE and therefore it can make full use of a Conditional GET header, enable caching of resources that can improve scalability of a system.

As highlighted the results from the benchmark testing. A RESTful service and in particular when using the messaging format of JSON, can be significantly lighter that its SOAP counterpart when traversing a network.

When the RESTful service comparison is shifted towards the XML messaging format, the results show that without caching, the choice of messaging format could be crucial as the XML format was out performed by both services, with it being 6.8% larger than the SOAP service and 54.8% larger than its JSON counterpart in relation to total data transported for the scenario.

Once the caching was enabled for the testing, the results show that there would be very little to choose between the two RESTful services, given that no change occurred to the underlying data. In the scenario, once caching was enabled, this then showed the total weight of data to be transported dropped to approximately 1% that of the SOAP service; which is unable to take advantage of the HTTP caching capabilities.

5 Conclusion

From the information put forward in this paper, as to which would be the more discerning choice, it would appear that due to the low impact of a RESTful service on the transport medium and with the utilisation of the HTTP headers enabling the use of Conditional GET caching, a RESTful service could be selected as the optimal choice and in particular the use of the JSON messaging format. Of course if fully standardised protocol usage was required and the transport impact was not a concern then the only way would be to utilise the WS-* standards.

5.1 Future Work

To further enhance this comparative analysis, we could extend the comparison to also include the development to the RESTful services, by means of also including OData alongside the JSON and XML RESTful services.

Then to be able to extend this discussion, a further comparative analysis of the WS-* stack and the RESTful services could be viewed to give a lower level comparison of both the conceptual and technological concepts that can be seen between the two styles. This then may be able to give a quantitative technical comparison basing on the architectural principles that can be drawn between them.

6 References

[1] Berners-Lee, T. . (2005) *Uniform Resource Identifier (URI): Generic Syntax* [Online]. Available from: http://labs.apache.org/webarch/uri/rfc/rfc3986.html [accessed: 17/12/11]

[2] Fielding, R.T. (2000) *Architectural Styles and the Design of Network-based Software*

Architectures. Doctoral dissertation, University of California, Irvine.

[3] Meier, J.D., Homer, A., Hill, D., Taylor,J., Bansode, P., Wall, L, Boucher, R. & Bogawat, A. (2008) *Application Architecture Guide 2.0, Patterns & Practices*

[4] Microsoft Corporation. (2009) *Fiddler2* [Online]. Available from: http://www.fiddler2.com/fiddler2/ [accessed: 01/10/12]

[5] Mulligan, G. & Graˇcanin, D. (2009) *A Comparison of SOAP and Rest Implementations of a Service Based interaction Independence Middleware Framework*

[6] Ort, E. (2005) *Service-Oriented Architecture and Web Services: Concepts, Technologies, and Tools*

[7] Richardson, L. & Ruby, S. (2007) RESTful Web Services

[8] Rodriguez, A. (2008) *RESTful Web Services: The basics*. [Online]. Available from: http://www.ibm.com/developerworks/webservices/library/ws-RESTful/index.html [accessed: 09/10/10]

[9] Stimulus Technology (2011) *StresStimulus* [Online]. Available from: http://stresstimulus.stimulustechnology.com/

[10] Tsai, W. T., Zhou, X. (2008) *SOA Simulation and Verification by Event-driven Policy Enforcement*. pp.165-172, 41st Annual Simulation Symposium 2008

Experimental Evaluation of Vector Bin Packing Algorithms on VM Consolidations in Cloud Data Centres

John Furlong, Lei Shi, Runxin Wang
TSSG, Waterford Institute of Technology, Ireland
Email: johnfurlong1@gmail.com; {lshi, rwang}@tssg.org

Abstract

As the proliferation of cloud data centres has increased rapidly (and continues to do so), so too has the power consumed in these enterprises. Strategies to reduce the power consumption can include variants of bin packing algorithms, which mean that physical machines (PMs) can be treated as bins and virtual machines (VMs) as items. This paper addresses the problem of which bin packing strategy to use in attempting to first consolidate VMs in a data centre (the aim of which is to reduce power consumption by powering off unused PMs), then going forward, which strategy to adopt when packing new VM requests to PMs as optimally as possible.

Keywords: Data Centre Optimisation, Vector Bin Packing, Green Computing

1 Introduction

The problem of how to minimise energy consumption in data centres has been an area of much research over the last number of years as the use and construction of new data centres has rapidly increased. Figures for 2010 show that of the total energy consumed in the world, between 1.1 and 1.5% was accounted for by data centres, with this figure rising to between 1.7 and 2.2% in the United States [1].

The move away from the dedicated server model to that of the cloud data centre helped reduce power consumption but also introduced a new set of problems regarding PM utilisation. This use of virtualisation technologies means that the well known optimisation problem commonly known as the ”bin packing problem” could be applied to cloud data centres to reduce power consumption. The VMs which are treated as items are packed onto PMs which act as the bins. These VMs and PMs can be sorted in different ways before the VMs are packed onto PMs in order to determine which sorting strategy leads to the most optimised data centre. In most cases the data centre is already running and therefore in an intermediate state. This is when VMs are dispersed randomly over PMs due to the dynamics of VM requests over time. VMs are chosen as candidates for migration from one PM to other PMs while the PMs which the VMs have been migrated from can be powered off, thus saving energy. To reduce the total power consumption the number of VM migrations (particularly live migrations in which VMs are migrated without stopping their services) must be reduced. Therefore VMs must be migrated from underutilised PMs to more utilised ones so that these underutilised PMs can then be powered off, but this must be done by also limiting the number of VM migrations.

In this paper, given the initial placement of the VMs and PMs, we show that by calculating the CPU and memory components of each and then sorting the VMs and PMs in different ways, that a significant energy saving may be possible in a cloud data centre. Next, a series of different algorithms are designed in order to consolidate the VMs onto as few PMs as possible but while also aiming to minimise the number of VM migrations necessary for this consolidation. A number of online algorithms are also designed for dynamic placement, aiming to place new VMs onto the data centres PMs as efficiently as possible. While similar work has been performed in this area, there are two main differences between previous work and our work. First is the scope; previous work was limited to testing only one or two sorting algorithms in order to gauge their consolidation effectiveness, this work tests multiple different sorting algorithms in order to identify which are the best performing algorithms in relation to various consolidation metrics. Secondly, previous research was only performed in relation to consolidation alone; a data centre in which VMs had become dispersed across PMs over time due to VM requests arriving and then dying out. The idea being that consolidation could be performed by the algorithms initially and then periodically, e.g. every hour. In this work we initially run the consolidation algorithms on a data centre such as this, however we also create a series of online algorithms which will use various strategies to pack new dynamic VM requests. This way we can consolidate the VMs in the first instance and then ensure that new VM requests are packed as optimally as possible.

The remainder of this paper is structured as follows, in §2 previous work in this area of research is discussed. Following on from that in §3 the idea behind this paper and the algorithms which were developed are explained in detail. In §4 we present experiments that evaluate the VM consolidation using our approaches for various VM request settings. Also presented is our interpretation of the results with evidence to support this interpretation. Finally we conclude the paper in §5.

2 Related Work

In relation to bin packing, Wilcox *et al.* [2] formulate a new algorithm called Reordering Grouping Genetic Algorithm (RGGA) which is first tested on conventional bin packing problems, before being applied to the specific bin packing problem of migrating and assigning VMs to PMs. Gupta *et al.* [3] view the problem of PM consolidation as falling into three categories: centralisation, physical consolidation and application integration, with the focus being on physical consolidation, which is closely related to the bin packing problem. Murtazaev & Oh [4] proclaim a PM consolidation algorithm for use in data centres which use virtualisation. The Sercon algorithm is specifically designed to be used with live migration in which applications are migrated from VM to VM without stopping their services. While these migrations are necessary for optimisation, the energy costs involved in a migration are large, therefore the Sercon algorithm will not only try to minimise the number of PMs used but also try to minimise the number of migrations also. Panigrahy *et al.* [5] take the approach of studying variants of the First-Fit Decreasing (FFD) algorithm in a heuristic approach to the Vector Bin Packing (VBP) problem.

In relation to data centre optimisation, Wang *et al.* [6] decide to focus on the bandwidth constraints of network devices, stating that it is difficult for the traditional schemes to make a ”reliable, deterministic estimate of bandwidth demand”. Data centres which are “power proportional”, i.e., they use power only in proportion to the load are described by Lin *et al.* [7]. There are three main challenges in relation to dynamically right-sizing the data centre, and the focus is on the third of these challenges which is the issue of choosing how many PMs to toggle into power-saving mode, and how to control these PMs and restarts.

An investigation into the power-aware provisioning of VMs for real-time services by Kim *et al.* [8] focuses on the cloud computing paradigm, the related “Anything as a Service” (XAAS) models, and the acknowledged Service Level Agreements (SLAs) between the cloud service providers and the customers which relates to Quality of Service (QoS) amongst other things. Beloglazov & Buyya [9] use “live-migration” and VM consolidation, but only focus on the QoS of such an approach even in heterogeneous infrastructure containing heterogeneous VMs. An attempt is made to solve both the traditional bin packing problem as well as the “intermediate state” data centre optimisation problem and four heuristics are proposed for choosing which VM to migrate. Beloglazov & Buyya [10] focus on the dynamic consolidation of VMs using live migration in which idle nodes are switched to sleep mode. Addressed are the online implications of the problem by conducting competitive analysis and by proving competitive ratios of the optimal online deterministic algorithms. These competitive ratios are proved for both single VM migration and dynamic VM consolidation problems.

Tarighi *et al.* [11] aim to find a more intelligent way to migrate VMs from underutilised PMs to more utilised ones. Their proposal is a new method which will migrate VMs between cluster nodes using the ”TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) algorithm”. TOPSIS is one of the most efficient Multi Criteria Decision Making techniques and is used to make more effective decisions over whole active PMs of the Cluster and find the most loaded PMs. Ajiro & Tanaka [12] propose using the FFD algorithm which is already widely used with regard to one and two dimensional bin packing problems but use it in conjunction with another algorithm called Least Loaded (LL) which is used for load balancing. This technique reduces the number of destination PMs by re-ordering the existing PMs and then retrying the packing procedure using either FFD or LL before a new PM is added.

3 Vector Bin Packing Algorithms

Given a data centre with an initial placement of VMs on PMs, and new VM requests arriving at the data centre at any time, the objective for the VM to PM placement problem is to minimise the active PM number, whilst subject to resource capacity constraints. A cloud data centre consisting of VMs hosted on PMs can be approached as a variant of the VBP problem, where each dimension is independent, and items are packed corner to corner diagonally, as long as the sum of the items (VMs) does not exceed the capacity of the bin in any dimension. As the bin-packing problem is known to be NP-hard and the problem outlined in this paper is a variant of the bin-packing problem, it too is NP-hard and as such optimal solutions cannot be guaranteed in polynomial time. However a number of well known heuristics such as FFD exist which provide suboptimal results [13].

We present six greedy VBP algorithms for PM consolidation, including PM Load (ServerLoad), Percentage Utilisation (PercentageUtil), Absolute PM Capacity (AbsoluteCapacity), FFD, FFD Sorted PM (FFDSorted), and FFD Residual Load (FFDResidual) algorithms. The initial algorithms were developed from well known heuristics such as FFD, and during the initial development and testing phase the variants of these greedy bin packing algorithms outlined above were proposed. The inputs to the algorithms are VMs V, PMs P in the data centre and the initial VM/PM mapping M. The outputs are an updated VM/PM mapping M and VM migration schedule S including VM ID, destination PM and source PM. In order to assess which PMs are underutilised, the scalar of every VM on a particular PM will be summed to determine the level of utilisation of the PM. The total magnitude of the VMs and PMs is calculated using the formula:

$$sqrt((\sum CPU)^2 + (\sum Memory)^2) \tag{1}$$

Once there is an insight into how densely each PM is populated, it is then a matter of going about assessing the relative size of each VM, which is the magnitude of each VM. Next, a consolidation algorithm can be applied in order to calculate the number of VM migrations needed to consolidate the PMs (each movement is virtual, i.e. no energy costs are involved, until the actual migration takes place). Once the migration strategy has been determined the process of VM migration can occur, freeing up PMs which can then be powered down, thus achieving an energy saving. An online packing strategy is then adopted in order to ensure that new VM requests are packed as optimally as possible with the minimum number of active PMs and high PM resource utilisation.

ServerLoad is specified in Alg. 1. A variable $FailedMig$ is initialized to 0, which counts the number of unsuccessful migration attempts. An unsuccessful migration attempt is when all the VMs on the least loaded PM are unable to be fully migrated to other PMs. The VM migration schedule S is initialized to $\emptyset$ in line 1. In lines 2-4, the occupied magnitude $O[i]$ of every PM i is calculated using equation 1. In line 5, PMs P are sorted in decreasing order based on the occupied magnitude. In lines 6-25, every PM will be processed to facilitate the migration of the VMs on the least loaded PM to the most loaded PMs. In line 7, copies of P, S and M are taken so in a situation where not all the VMs on the least loaded PM are fully migrated, the state of P, S and M can be restored. In line 8 the VMs $V[i]$ on each PM are sorted in decreasing order so that the VM with the largest magnitude is the first candidate for migration. In lines 9-17, VM $V[i][j]$ which has the largest magnitude on PM i, (which is the PM with the lowest occupied magnitude) is the first candidate for migration. This VM is attempted to be placed on the PM with the highest occupied magnitude, if unsuccessful the PM with the second highest occupied magnitude is tried,

Algorithm 1 ServerLoad Algorithm

```
Input: $V$, $P$, $M$
Output: $S$, $M$
1:  $FailedMig = 0$, $S = \emptyset$
2:  for $i \leftarrow 1$ to $length[P]$ do
3:      Calculate the occupied magnitude O[i]
4:  end for
5:  Sort $P$ in descending order of occupied magnitude
6:  for $i \leftarrow length[P]$ to 1 do
7:      Backup $P$, $S$, $M$
8:      Sort VMs $V[i]$ in PM $i$ in decreasing order of magnitude
9:      for $j \leftarrow 1$ to $length[V[i]]$ do
10:         for $k \leftarrow 1$ to $i$ do
11:             if $V[i][j]$ fit into PM $k$ then
12:                 $S \leftarrow S \cup \{V[k][j], i, k\}$
13:                 update occupied magnitude $O[k]$ of PM $k$
14:                 Sort $P$ in descending order of occupied magnitude
15:             end if
16:         end for
17:     end for
18:     if VMs$V[i]$ are not fully migrated then
19:         Restore $P$, $S$, $M$
20:         $FailedMig$++;
21:     end if
22:     if $FailedMig \geq FailedMigLim$ then
23:         Report error and return
24:     end if
25: end for
```

and so on. If there are sufficient resources for this VM to be placed, then S is updated with that VMs identifier $V[i][j]$, the source PM i, and the destination PM k. In addition, the occupied magnitude $O[k]$ of the destination PM k is also updated to reflect the placement of the new VM $V[i][j]$, and PMs are again sorted in descending order based on occupied magnitude. This process repeats until all the VMs on the least loaded PM are migrated, if all the VMs are migrated successfully, then the process moves on and the next least loaded PM is selected, and so on. If not all the VMs on the least loaded PM are fully migrated then we will need to revert back to the previous states of P, S and M and the number of $FailedMig$ is incremented. If $FailedMig$ is greater or equal to a predefined value $FailedMigLim$ then the progress is terminated, as shown in lines 18-24.

The remaining algorithms all operate in a similar fashion, with only variations mentioned below:

- PercentageUtil: The PMs P are sorted in decreasing order based on the ratio of absolute capacity of the PM against the total magnitude of any VMs hosted on it.
- AbsoluteCapacity: The PMs P are sorted by absolute PM magnitude in decreasing order.
- FFD: VMs V are sorted in decreasing order and we suppose initially all the PMs P are empty. Each VM in turn is then attempted to be placed on the first PM that will accommodate it, where PMs are sorted in terms of their PM IDs. The initial mapping and the final mapping created by FFD are then compared. If a VM is mapped to the same PM on both mappings then no migration is necessary, if not then the VM ID, the original PM ID and the new PM ID are added to the migration schedule S.

- FFDSorted: operates the same way as FFD except the PMs are also sorted in decreasing order based on their absolute capacity.
- FFDResidual: operates the same way as FFD except the PMs are also sorted in decreasing order by their absolute residual capacity. Initially, the absolute residual capacity for a PM is its total magnitude.
- Online Algorithms : New VM requests arrive and are attempted to be placed in turn based on whatever way the particular algorithm sorts the PMs, (which are defined by their consolidation algorithm counterparts). If a VM can be accommodated on a PM, the VM/PM mapping is updated to reflect the new VM placement. If the VM cannot be placed on any of the PMs then the request is rejected. Each VM has a lifespan, when its lifespan has expired, the VM is removed from its hosted PM and the consolidation algorithms are applied to consolidate the PMs.

4 Performance Evaluation

4.1 Experiment Setup

The initial placement is created with 500 randomly generated VMs from the four VM types as per the Amazon EC2 instances [14]. Each of these VM types has a different memory and CPU capacity which is used to calculate its magnitude. The total magnitude of all 500 VMs is calculated, then tripled. This is done to ensure that the dispersed state of a data centre containing underutilised PMs (with an average server utilisation of between 25% and 33% as per other research in the area [4]) is represented. Next a number of PMs are generated at random from four different PM types, until their available magnitude is greater or equal to the tripled total VM magnitude. The four PM types are chosen to be an order of eight times larger than each of the four VM types as per other research in the area [4].

VMs are then placed on PMs in a random mapping to recreate the dispersed state of a data centre which may occur over time as new VM requests arrive and old VM requests expire. Once the initial VM/PM mapping is in place, each of the six consolidation algorithms are applied. Each of these algorithms receive the same initial VM/PM mapping as an input so their performance can be accurately measured. Each algorithm will devise a new VM/PM mapping in their own way as they attempt to migrate VMs from one set of PMs to another. This balance between consolidation and VM migrations is measured using the migration efficiency metric. The migration efficiency is the ratio of the number of PMs released to total migrations.

A number of other metrics also measure various facets of the algorithms performance such as number of migrations, number of PMs used and PM utilisation percentage. Number of migrations refers to the number of VMs which need to be migrated in order to achieve a consolidated data centre environment, this number should ideally be as small as possible. Number of PMs refers to the number of active PMs after the consolidation has occurred, this figure should ideally be as small as possible. PM utilisation percentage is a measure of the ratio of used magnitude on all the PMs to the total magnitude of all the PMs. Ideally we would like this figure to be as close to 100% as possible.

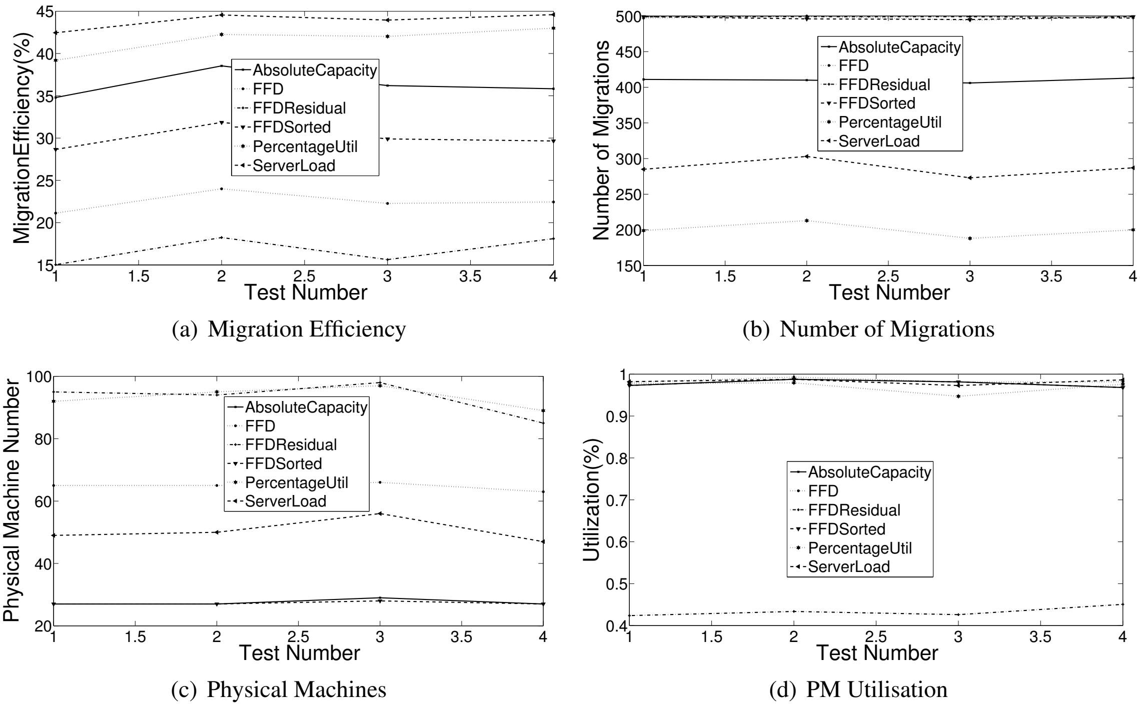
For the dynamic placement, the test regime was run four times which meant four different sets of inputs. For each test case, we start with a data centre with initial placement from time 0, and the number of VM requests is modelled as a Poisson process with average $\lambda = 20$. The VM types are randomly generated from the four different VM types. Each VM has two randomly generated numbers to represent the starting time at which the VM request arrives and the VMs lifespan. The random "seed state" is set so that the list of new VM requests generated is the same for every online algorithm so that each can be compared accurately.

4.2 Experiment Results

In the consolidation phase, the overall best performing algorithm was ServerLoad. This algorithm was the best performing by a small but clear margin when it came to migration efficiency as shown in Fig 1(a).

It was the second best performing in relation to both the number of migrations in Fig 1(b) and the number of physical machines used in Fig 1(c). Finally ServerLoad was one of a number of algorithms which had almost 100% server utilisation as shown in Fig 1(d), and therefore was one of the best performing algorithms in this respect too. The next best performing algorithm was PercentageUtil, this was a close second behind ServerLoad regarding migration efficiency as shown in Fig 1(a), the best in relation to number of migrations in Fig 1(b) and along with a group of algorithms, one of the best when it came to server utilisation with almost 100% in Fig 1(d). However regarding the number of physical machines, PercentageUtil was actually the worst performing as shown in Fig 1(c).

The third best performing algorithm for migration efficiency was AbsoluteCapacity as shown in Fig 1(a). This algorithm was also the third best regarding the number of migrations as shown in Fig 1(b) ahead of the three FFD based algorithms. In relation to number of physical machines this algorithm was actually the best performing in Fig 1(c), and was one of a number of algorithms which were the best when it came to server utilisation as shown in Fig 1(d).



(a) Migration Efficiency

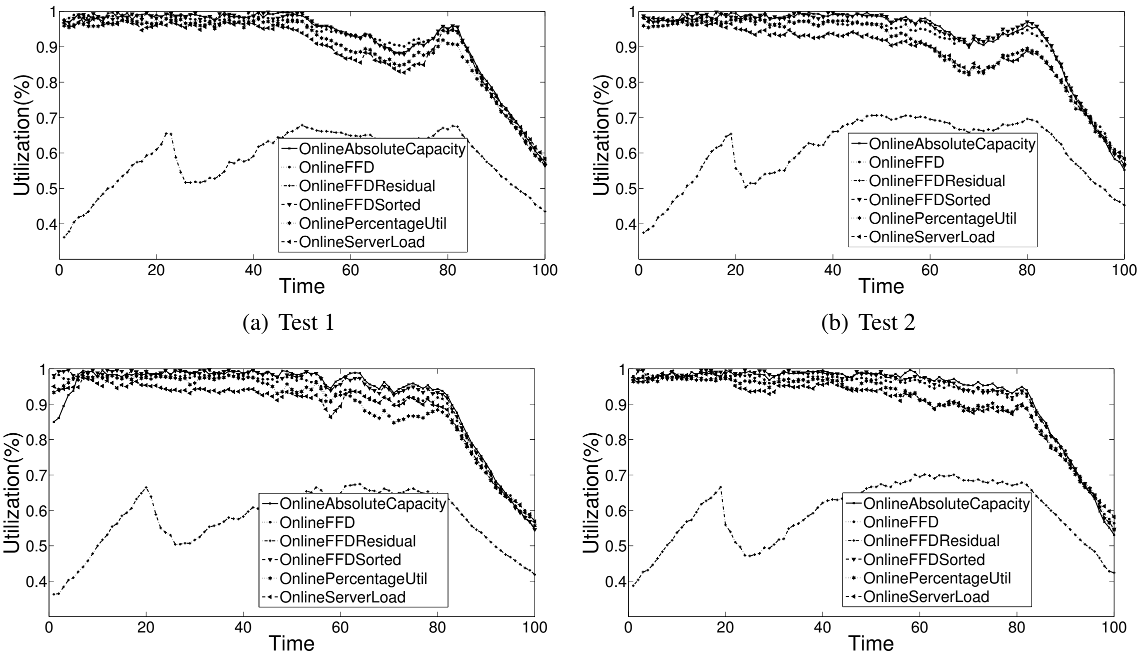
(b) Number of Migrations

(c) Physical Machines

(d) PM Utilisation

Figure 1: Six consolidation algorithms and their performance on various metrics.

The next best performing algorithm (and best performing of the FFD based algorithms) was FFDSorted. This was the fourth best performing in relation to migration efficiency as shown in Fig 1(a). In fact if the number of migrations had been approximately 25% lower, FFDSorted would have actually been the best performing when it came to migration efficiency, it is the fact that the number of migrations needed was almost the maximum possible that reduced the performance of FFDSorted. However, all of the FFD based algorithms (FFD, FFDSorted, FFDResidual) have similarly poor results in relation to number of migrations as shown in Fig 1(b), and FFDSorted is no different. The poor performance of the FFD based algorithms is perhaps unsurprising given that they operate by comparing the initial random VM/PM mapping with an FFD generated mapping. In relation to the number of physical machines used, FFDSorted was actually the best performing, slightly ahead of AbsoluteCapacity in Fig 1(c). Regarding server utilisation, FFDSorted was among a number of algorithms which had just under total utilisation as shown in Fig 1(d). The second worst performing algorithm was FFD. This was the second worst performing when it came to migration efficiency as shown in Fig 1(a). As previously stated, FFD is one of the joint worst performers in relation to the number of migrations in Fig 1(b) and was the fourth worst regarding the number of physical machines in Fig 1(c). FFD was among a number of algorithms which



(a) Test 1

(b) Test 2

(c) Test 3

(d) Test 4

Figure 2: PM Utilisation Vs Time for online placement of the six packing algorithms.

had just under total utilisation as shown in Fig 1(d).

By far the worst performing algorithm across a number of metrics was FFDResidual. Regarding migration efficiency FFDResidual was clearly the worst performer as shown in Fig 1(a). As an FFD based algorithm, FFDResidual is one of the joint worst performers in relation to the number of migrations in Fig 1(b) and was also the joint worst performer along with PercentateUtil when it came to the number of physical machines used in Fig 1(c). Even in relation to server utilisation, while the other five algorithms had an almost identical level of utilisation of just under 100%, FFDResidual only had a server utilisation of around 40%. The poor performance of FFDResidual can be explained simply due to how the algorithm operates; VMs are placed onto the PM with the most spare capacity each time. This approach means that VMs are dispersed across a wide range of PMs with no PM very consolidated. The time complexity of all the algorithms is between $(O(log(n^4))$ in the case of more complex algorithms such as ServerLoad and $(O(log(n^2))$ in the case of algorithms such as FFD.

Regarding the online placement, as shown in Fig 2, there were small differences in how well most of the algorithms packed dynamic VM requests in terms of PM utilisation. However, it is clear that online FFDResidual was the worst performing algorithm. While this algorithm is the least effective for this particular application, there may exist other applications requiring this feature, such as load-balancing. Although the online algorithms are based on the sorting and packing strategies of their consolidation counterparts, it is important to note that they are independent of each other, so the most effective consolidation algorithm may be unrelated to the most effective dynamic placement algorithm, and both these algorithms can work together.

5 Conclusion and Future Work

We designed and evaluated a number of bin packing algorithms using different sorting strategies to determine firstly which one is most effective at consolidating VMs in data centre enabling unused PMs to be turned off. Futhermore, various strategies were developed to pack dynamic VM requests in the most efficient way. Each of the six algorithms offered some improvement in the number of PMs which could be powered off due to consolidation, even when the number of VM migrations needed for this consolidation was also taken into account. There are several interesting future research directions motivated by our work. Firstly the initial data centre placement instead of being generated randomly could perhaps take the guidance of a combinatorial auction [15] similar to Amazon's EC2 spot instances [14]. This would ensure that the initial data centre placement bears a more realistic resemblance to a real world data centre. Second, a number of constraints relating to fault tolerance or security could also be considered. Last but not least, PM failures should be considered for the dynamic VM placement.

References

[1] J. Koomey, "Growth in data center electricity use 2005 to 2010," *The New York Times*, 2011. [Online]. Available: http://www.analyticspress.com/datacenters.html

[2] D. Wilcox, A. W. McNabb, and K. D. Seppi, "Solving virtual machine packing with a Reordering Grouping Genetic Algorithm," in *IEEE Congress on Evolutionary Computation*, 2011, pp. 362–369.

[3] R. Gupta, S. K. Bose, S. Sundarrajan, M. Chebiyam, and A. Chakrabarti, "A two stage heuristic algorithm for solving the server consolidation problem with item-item and bin-item incompatibility constraints," in *Proc. IEEE SCC*, Washington, USA, 2008, pp. 39–46.

[4] A. Murtazaev and S. Oh, "Sercon : Server Consolidation Algorithm using Live Migration of Virtual Machines for Green Computing," *Iete Technical Review*, vol. 28, no. 3, pp. 212–231, 2011.

[5] R. Panigrahy, K. Talwar, L. Uyeda, and U. Wieder, "Heuristics for Vector Bin Packing," *research-microsoftcom*, 2011.

[6] M. Wang, X. Meng, and L. Zhang, "Consolidating virtual machines with dynamic bandwidth demand in data centers," in *INFOCOM*. IEEE, 2011, pp. 71–75.

[7] M. Lin, A. Wierman, and L. L. H. Andrew, "Dynamic right-sizing for power-proportional data centers," *Queue*, pp. 1098–1106, 2011.

[8] K. H. Kim, A. Beloglazov, and R. Buyya, "Power-aware provisioning of virtual machines for real-time Cloud services," *Concurrency and Computation: Practice and Experience*, vol. 23, no. 13, pp. 1491–1505, 2011.

[9] A. Beloglazov and R. Buyya, "Energy Efficient Allocation of Virtual Machines in Cloud Data Centers," in *IEEE CCGrid*, 2010, pp. 577–578.

[10] ——, "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, 2012.

[11] M. Tarighi, S. A. Motamedi, and S. Sharifian, "A new model for virtual machine migration in virtualized cluster server based on Fuzzy Decision Making," *CoRR*, vol. abs/1002.3329, 2010.

[12] Y. Ajiro and A. Tanaka, "Improving Packing Algorithms for Server Consolidation," in *Int. CMG Conference*. Computer Measurement Group, 2007, pp. 399–406.

[13] E. G. Coffman, M. R. Garey, and D. S. Johnson, "Approximation algorithms for bin packing: a survey," in *Approximation algorithms for NP-hard problems*. Boston, MA, USA: PWS Publishing Co., 1997, pp. 46–93.

[14] Amazon. Amazon ec2 instance types. [Online]. Available: http://aws.amazon.com/ec2/instance-types/

[15] S. Zaman and D. Grosu, "Combinatorial auction-based allocation of virtual machine instances in clouds," in *IEEE CloudCom*, Washington, DC, USA, 2010, pp. 127–134.

Protecting Organizational Data Confidentiality in the Cloud using a High-Performance Anonymization Engine

Vanessa Ayala-Rivera [1], Dawid Nowak [1], Patrick McDonagh [2]

[1] Lero@UCD, School of Computer Science and Informatics, University College Dublin.
vanessa.ayala-rivera@ucdconnect.ie; dawid.nowak@ucd.ie
[2] Lero@DCU, School of Electronic Engineering, Dublin City University.
patrick.mcdonagh@dcu.ie

Abstract

Data security remains a top concern for the adoption of cloud-based delivery models, especially in the case of the Software as a Service (SaaS). This concern is primarily caused due to the lack of transparency on how customer data is managed. Clients depend on the security measures implemented by the service providers to keep their information protected. However, not many practical solutions exist to protect data from malicious insiders working for the cloud providers, a factor that represents a high potential for data breaches.

This paper presents the High-Performance Anonymization Engine (HPAE), an approach to allow companies to protect their sensitive information from SaaS providers in a public cloud. This approach uses data anonymization to prevent the exposure of sensitive data in its original form, thus reducing the risk for misuses of customer information. This work involved the implementation of a prototype and an experimental validation phase, which assessed the performance of the HPAE in the context of a cloud-based log management service. The results showed that the architecture of the HPAE is a practical solution and can efficiently handle large volumes of data.

Keywords: Cloud Computing, SaaS, Data Confidentiality, Data Anonymization, Performance.

1 Introduction

In recent years, cloud computing [1] has become a popular business model and a promising technology paradigm for delivering IT services. This model offers multiple benefits to the diverse interests of customers and providers. For example, small and medium-sized enterprises (SMEs) find the cloud a potential area to expand their market and reach new clients. Consequently, enterprises from various domains are already using or planning to implement some type of cloud service in the near future [2, 3].

Despite these advantages, a large number of companies are still reluctant to use the cloud due to the associated risks of adoption [4]. According to the work presented in [5, 6, 7, 8], security continues to be the primary obstacle preventing the adoption of cloud services. In a time where data is a critical asset for many firms, data protection has become one of the top security concerns [9]. Therefore, companies are wary of losing control over their sensitive data by placing it on platforms that they do not manage.

Under these circumstances, customers depend on the cloud providers to have the appropriate security measures in place to protect the data. Nevertheless, the fact that customer data needs to be in a simple text format to be used, raises privacy concerns about potential attacks from malicious insiders working for providers. This threat is amplified due to the lack of transparency about providers' processes [10]. This situation is more concerning in the SaaS delivery model, because it offers the highest level of abstraction, hence offering the least visibility on how data is managed.

According to the 2009 Data Breach Study performed by the Ponemon Institute [11], over 44% of cases related to general data breaches in the industry were caused by third party contractors, partners and outsourced vendors. This and other similar worrying statistics [12] confirm companies' concerns about data security. These statistics should also encourage customers of cloud-based services to implement their own internal controls to compensate for the potential security deficiencies of providers.

Our work aims to facilitate customers in the implementation of their own methods to protect their critical information in-house, before uploading their data to the cloud. In this paper, we present the High-Performance

Anonymization Engine (HPAE), which is our proposed approach to protect the confidentiality of customers' data when using SaaS applications. As part of our work, we have implemented a prototype application and conducted a set of experiments to validate the feasibility of our approach.

The structure of this paper is as follows. Section 2 provides a review of the related work. Section 3 describes the proposed approach, including details of the implementation. Section 4 presents the results of the experimental validation. Finally, Section 5 presents the conclusions and future work.

2 Related Work

Data confidentiality has been an active research area in recent times as it remains a top concern for adoption of cloud computing model. As a result, many different approaches have been proposed to ensure data security in the cloud. One proposed solution is to simply avoid external clouds and build in-house private clouds. In this scheme, companies try to retain the advantages of the cloud model by employing private/hybrid cloud initiatives, hence avoiding the issues of public clouds [8]. However, this approach is expensive and not affordable for most companies.

Another alternative for data protection is to use traditional cryptography techniques to encrypt all cloud data. While this technique might be a good solution to protect data when it is transmitted or stored at the vendor side, it is not appropriate for data which is used for computation. The problem is that this technique highly restricts further data use, such as searching and indexing. Some state-of-the-art cryptography works have offered more versatile encryption schemes that allow operations upon and computation on the ciphertext [13, 14, 15]. However they are still too slow to be viable for real-world applications. Another encryption approach is Silverline [16], which identifies and encrypts all functionally encryptable data (any sensitive data that can be encrypted without limiting the functionality of the application in the cloud). However, the applicability of this approach is also limited because it assumes that web applications do not require access to raw data, which is rarely the case.

Our approach is closely related to the work described in [17], in the context of using data obfuscation to protect sensitive attributes. However, their solution requires cooperation from the service providers to implement logic on their side, situation which is not always feasible. Another approach related to our work is presented in [18], which also aimed to protect data from cloud service providers. Here, the authors describe three conditions to prevent that users' confidential information be collected by service providers. Firstly, separate software and infrastructure service providers. Secondly, hiding information about the owners of the data, and finally, the use of data obfuscation. Nevertheless, this flexibility is not always possible as it is often the case that the provider offering the software manages the infrastructure or platform service as well, so the user has no control over this.

3 Proposed Approach

The context of our approach is shown in Figure 1. In most companies, different data sources exist within the secure boundaries of the enterprise intranet. However, whenever an interaction occurs with a SaaS application, the company's data might leave the security of the intranet and be transferred to the SaaS provider. When this scenario arises, a preferable situation is that the information could be protected before being sent. For this purpose, companies can implement their own methods to protect their critical information in-house, before uploading their data to the cloud. This will keep the data safe from potential misuse by the service provider, while still retaining utility to be processed and analyzed. One technique that could be used for this purpose is data anonymization, which is the process of altering the original data in such way that it is difficult to infer anything private about the entities represented. This simultaneously limits the loss of information such that data is still meaningful permitting its analysis. Similarly, in some cases, users may need to access their original data. Once data is processed in the cloud, the output can then be returned to the enterprise secure intranet and a reversibility mechanism can be used to reveal the original values from the anonymized data. Our approach, the HPAE, aims to fulfil those responsibilities in Figure 1. The following sections describe in detail the components of our proposed architecture and implementation.

Finally, this work has the following goals:

- The development of an approach employing anonymization to protect confidential data on a public cloud. The objective is to protect customers' sensitive data, using the HPAE, from SaaS providers when data is processed in the cloud. Moreover, as the concept of confidentiality varies among users, our solution aims to be flexible enough to allow users to configure their privacy policies.
- The design of an architecture that efficiently handles large volumes of data offering high-throughput and fast processing. Performance is a determining factor in the adoption of a new approach, thus we aim to

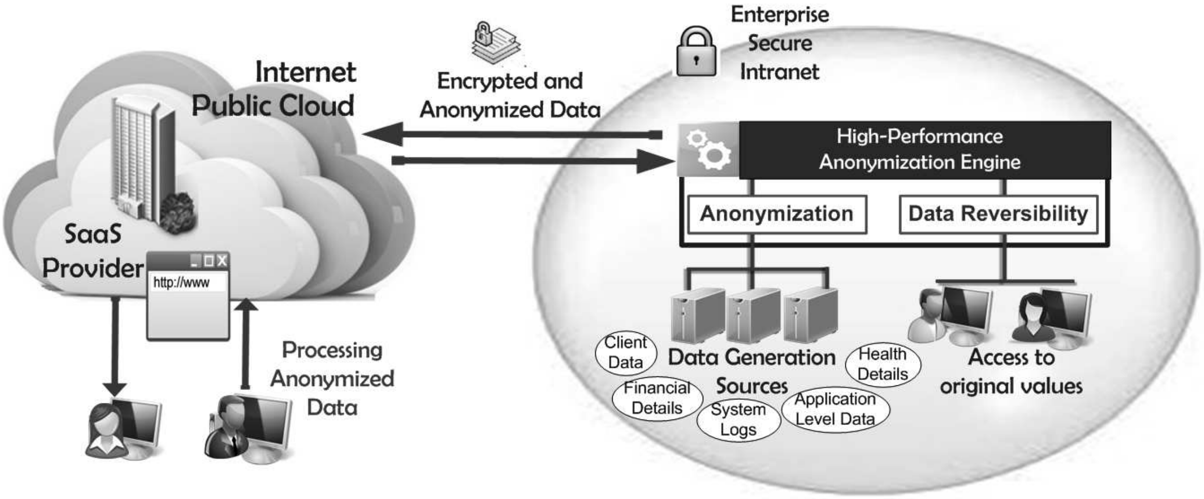


Figure 1: Contextual View of the HPAE

provide a solution that offers good performance (in terms of throughput) such that it is practical and allows users to anonymize data on-the-fly. Furthermore, the design should be modular and extensible to facilitate the accommodation of new components, such as new anonymization techniques, new input/output types, more efficient libraries/data structures etc.

- The development of a prototype tool in Java to demonstrate our approach and measure the performance of our architecture. To facilitate ease of integration of the HPAE with existing systems inside organizations, our Java implementation will provide support for various types of input data sources as well as various types of output destinations.

3.1 HPAE Architecture

This section describes the different components in the architecture of our prototype. The HPAE is the core component and it is composed by one-to-many *Data Processor* (DP) threads; although only a single DP is shown here, there may be as many DPs as input types as denoted in Figure 2. There are four stages in our approach: *Configuration, Reading, Anonymization* and *Writing*. The prototype tool for the HPAE is still currently being developed; more details about the implementation details are provided in Section 3.1.5.

3.1.1 Configuration

In this stage two files are configured: the *Engine Descriptor* and the *Rules Descriptor*.

The *Engine Descriptor* is an XML file containing the configuration parameters to initialize the HPAE and set values for the DPs. Among others, some of the parameters found in this file are: input sources/output destinations for DPs, the *Rules Descriptor* file for each DP and other technical properties like buffer size and the number of Event Handler threads to process the data.

The *Rules Descriptor* is an XML file where the rules to identify sensitive information are defined (i.e. IP addresses, emails, etc.). The rules are defined in the form of patterns that can be position based, expressions or occurrences of specific strings. One file can be configured for each input source.

Once the descriptor files have been configured, they are validated. The *Configuration Parser* is in charge of parsing the XML files and validating that parameter values are correct (i.e. positive numeric values, duplicate values, etc.). The output of this process is either; presenting a list of configuration errors that need to be fixed to the user, or passing the list of DPs to be run to the HPAE. The configuration of the descriptor files is currently performed manually as a *GUI* has yet not been implemented.

3.1.2 Reading

The components used in the reading process will correspond to the selected type of input. In order to minimize the effort required by organizations in making changes to their current systems (which can have different technologies implemented), the HPAE supports the most common input sources and output destinations. The HPAE reads and

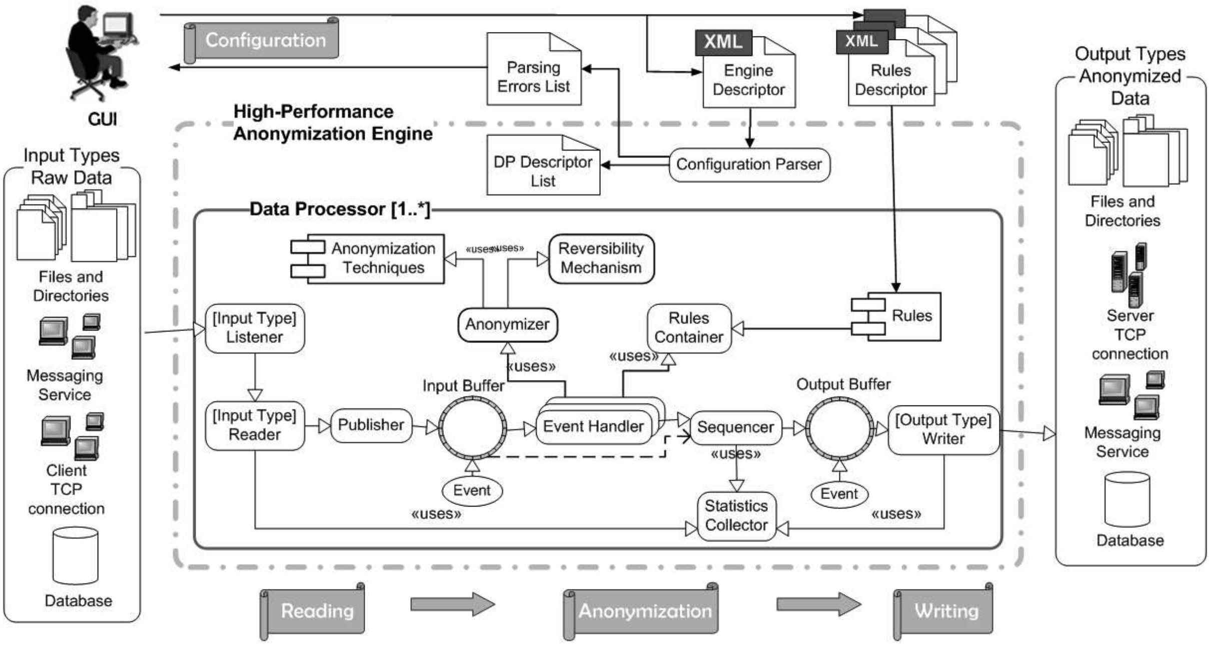


Figure 2: Technical View of the HPAE

writes data from/to databases, files, directories, TCP connections and messaging services like queues or topics. Support for the database type has yet not been implemented and remains as future work.

To start the data processing, the corresponding *Listener* waits for data to become available. For example, in the case of TCP connections, the *Listener* will wait until a connection is made to the specified port number. Once a client connection is accepted, the corresponding *Reader* will start receiving data from the socket. The received payload is encapsulated in an *Event* object, which can be a line of text in a file, a message from a queue/topic, a tuple from a database, etc. These *Events* are sent to the *Publisher*, which claims the next available slot in the *Input Buffer* and adds the entry. Buffers are circular data structures used to exchange the data between the different processing stages. These stages are asynchronously processed, meaning, one stage (i.e. Anonymization) does not need to wait until the previous one is fully finished (i.e. a file fully read) to start its processing.

3.1.3 Anonymization

The *Event Handlers* are the threads that fetch the *Events* from the *Input Buffer* and process them. The logic to publish entries to the buffer and retrieve them is based on the *Disruptor* pattern [19], explained in more detail in Section 3.1.5. Each *Event Handler* processes one *Event* from the buffer. To apply the anonymization, the *Event Handlers* use the *Rules Container* and the *Anonymizer*. The *Rules Container* has the patterns that are used to identify sensitive information. The *Anonymizer* contains the set of anonymization techniques to be applied. The *Event Handlers* perform the pattern matching; if a match occurs, the anonymization technique for that attribute is applied. Our prototype uses data substitution, extracting the sensitive attribute and replacing it with a new token in the *Event*.

As some anonymization techniques are not reversible, it is desired to have mechanisms that allow re-identification of the data regardless of the selected technique. In some scenarios, customers might require the original information, for example to perform root cause analysis of an incident investigation. The HPAE aims to provide various forms of *Reversibility Mechanism* to retrieve the original value such as keeping a translation table that tracks all the transformations done to the data or building dictionaries on-the-fly. This latter could also provide a consistent anonymization (using the same value mapping) across multiple data streams, which can be useful to correlate information from several sources. Currently, the *Reversibility Mechanism* remains part of our future work.

Since the HPAE can have multiple *Event Handlers* working in parallel, there is no guarantee on the order they will finish processing an *Event*. Thus, to ensure that the *Events* exit in the same order as they were received, the *Sequencer* is used. This component iterates the *Input Buffer* and retrieves the *Events* (once they have been anonymized) to publish them in the correct order to the *Output Buffer*.

3.1.4 Writing

Based on the output type defined for the DP, the corresponding *Writer* is created. This component retrieves the anonymized *Events* from the *Output Buffer* and sends them to the specified destination. As an optional step, if the user indicated in the *Engine Descriptor* file to gather performance statistics for the DP such as throughput, these are calculated by the *Statistics Collector*. In the current implementation, statistics are collected and written to a CSV file for analysis.

3.1.5 Implementation Details

Our prototype was implemented in Java because of its object-oriented nature, built-in support for multi-threading, portability and the practical benefits it possesses such as vast amounts of libraries. Furthermore, Java is one of the most widely used languages in enterprise applications [20], which can facilitate the adoption of our solution. To achieve high performance, the architecture of the HPAE was designed to execute three main phases (depicted in Figure 2) as independent processes (in separate threads). Given these characteristics, we investigated the use of different high performance libraries that we could apply to our design. In the end, we decided to use the *Disruptor* framework created by LMAX Exchange [19]. This framework offers data structures and a pattern of use (composed of producers, a ring buffer and consumers) that deal efficiently with concurrent programming, improving time and memory allocation compared to regular Java queue implementations that tend to suffer from write contentions. For example, in a pipeline configuration, instead of using a queue to exchange information between stages, which introduces latency, *Disruptor* uses a single lock-free data structure that deals with concurrent access. This framework fits well with our design as it is intended for an asynchronous event processing architecture like ours. Furthermore, our design is flexible enough to accommodate other types of data structures if needed.

3.1.6 HPAE for Log Management Service in the Cloud

One of the areas that suits the cloud model, in terms of configurable resource provisioning, is log management (LM) [21]. This kind of service successfully overcomes the main challenges faced by the SMBs: balancing the limited amount of in-house resources (people and infrastructure) against the ever-increasing supply of log data. However, log data contains aspects that can compromise the safety of the enterprise, thus it is recommended to protect confidential data when these logs are transmitted to a third-party such as cloud service providers.

The inherently sensitive nature of logs and the steady growth of data make LM an ideal application scenario for our approach, where our prototype is aimed at anonymizing sensitive data in the logs. For experimental purposes, we defined the IP address attribute as the pattern of sensitive data in the *Rules Descriptor* file, as this is the most commonly anonymized field in security and network relevant logs. For our initial prototype, the selected anonymization technique was data substitution, which consisted of matching the desired pattern in the logs and replace it with a token, which in our case was a Base64 encoded version of the original value. This technique was selected because it is fast in terms of processing time. This characteristic allowed us to minimize the overhead caused by the anonymization process and assess exclusively the feasibility of our approach.

4 Experimental Evaluation

This section describes the experiments aimed to assess the performance of the HPAE. The experiments were performed on a machine running 64-bit Windows 7, with an Intel Core i7 processor (4 cores / 8 threads) at 2.0 GHz clock speed with 6 GB RAM using Java HotSpot 1.6.0_31. The first experiment validated the selection of the *Disruptor* framework for our implementation, while the second experiment validated the performance of the HPAE prototype.

4.1 Experiment 1: Disruptor Performance Testing

The objective here was to validate the efficiency of *Disruptor* against Java's *ArrayBlockingQueue* (identified by LMAX as the Java bounded queue with the best performance [22]) to determine if *Disruptor* could be a better structure for data transfer than the traditional queues. This test, conducted in two phases, involved the assessment of two different metrics: throughput and latency. All the tests used in this experiment were taken from *Disruptor* version 2.8 open source project available at [19].

Phase 1 - Throughput Performance Testing.

The aim of this test was to compare the throughput offered by *Disruptor* and *ArrayBlockingQueue* when given three different configurations. The topologies that were selected for this experiment were the ones found within

Table 1: Throughput Comparison between Disruptor and ArrayBlockingQueue

Configuration	**Data Structure**	**Throughput (Operations per Second)**		
		Best case	**Average case**	**Worst case**
Unicast (1 P - 1 C)	ArrayBlockingQueue	3,424,305	3,340,106	3,256,586
	Disruptor	51,921,079	47,324,781	45,641,259
Multicast (1 P - 3 C)	ArrayBlockingQueue	614,352	592,084	564,174
	Disruptor	64,432,989	33,724,278	23,036,166
Pipeline (1 P - 3 C in stages)	ArrayBlockingQueue	1,757,716	1,737,686	1,721,763
	Disruptor	40,617,384	30,688,199	20,777,062

Table 2: Latency Comparison between Disruptor and ArrayBlockingQueue

Data Structure	**Latency (Nanoseconds)**		
	Mean	**Max**	**99% less than**
ArrayBlockingQueue	6,750	6,142,182	8,192
Disruptor	25	397,500	2

our architecture: Unicast, which consists of one producer (P) to one consumer (C); Multicast, which consists of one producer to multiple consumers (3 in our test) and Pipeline, which consists of a chain of one producer to multiple consumers where each consumer depends on the output of the previous consumer.

Table 1 shows the throughput results, in terms of operations per second, for 5 runs processing 100 million messages. The results reported in this table are for the best, worst and average case scenarios. These results demonstrate how *Disruptor* has a greater throughput performance than *ArrayBlockingQueue* for all cases. For example in the average case, *Disruptor* is 14 times better in the Unicast configuration; similarly, *Disruptor* is 56 times better in the Multicast configuration and 17 times better in the Pipeline one.

Phase 2 - Latency Performance Testing.

The aim of this testing was to compare the latency introduced by *Disruptor* and *ArrayBlockingQueue* for a configuration of three stage pipeline. Entry events were generated and injected in intervals of 1 microsecond to prevent saturation, repeating this process 50 million times. Table 2 shows the latency performance results for *Disruptor* and *ArrayBlockingQueue*. We can observe that *Disruptor* outperforms the queue implementation in all comparisons. For example, in the 99% of observations the latency of *Disruptor* was minimal (only 2 nanoseconds), while *ArrayBlockingQueue* took more than 8,000 nanoseconds.

To summarize the results of Experiment #1, they demonstrated that *Disruptor* offers better throughput and lower latency than the best Java Queue class (*ArrayBlockingQueue*). Therefore, it was decided to use *Disruptor* as part of our prototype implementation.

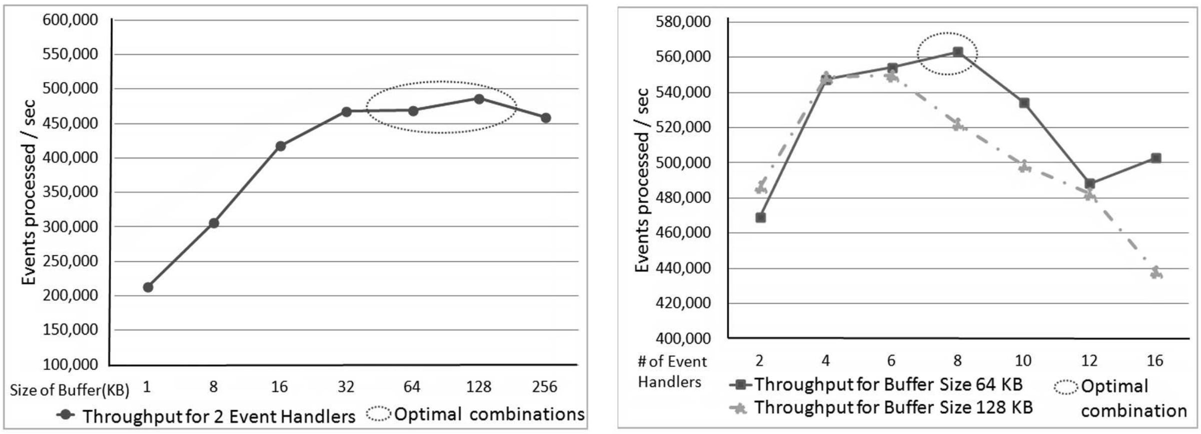
4.2 Experiment 2: HPAE Performance Testing

The objective of this experiment was to assess the efficiency of our architecture and implementation. We conducted a series of performance tests divided in two phases: The first phase focused on finding the optimal set of configuration parameters, while the second phase focused on the evaluation of the throughput using that configuration. For these tests, the HPAE was set up to run one socket DP and receive data from a client TCP connection in an isolated network. A second laptop was used for the client connection which has the same characteristics of the machine running the HPAE.

Phase 1 - Find the optimal parameters for HPAE: Size of Buffers and Number of Event Handlers.

The goal of this phase was to find the optimal combination of parameters for the HPAE, which maximized the throughput, measured in events processed per second (eps).

Figure 3 shows the results for the throughput performance tests for 5 runs processing 70 million log events for different combinations of parameters. In Figure 3 (a) we searched for the optimal size for the buffers, thus this was the variant parameter and the number of *Event Handlers* was set to 2. Given a uniform workload, we can see that when the size of the buffer is incremented, the throughput also increases until it reaches the maximum point at 128 KB. After this, the HPAE reaches a steady state (experiencing queuing delay in a saturated buffer) where the throughput slightly decreases. As a result of this experiment, we took sizes of 64 and 128 KB as the optimal values. Using these two values as constants, we then investigated the optimal number of *Event Handlers*, thus this was the variant factor in a second experiment. In Figure 3 (b) we can observe that using the buffer size of 64 KB, the throughput rate increases as the number of *Event Handlers* is augmented until reaching the maximum point



(a) (b)

Figure 3: Throughput Performance Testing: (a) Finding the optimal buffer size; (b) Finding the optimal number of Event Handlers.

Table 3: Results for the HPAE performance testing

Measure	Average	Max
Throughput (eps)	429,514	563,066
CPU usage (%)	33.44	34.5
Memory usage (MB)	85	102

(around 560,000 eps) at 8 *Event Handlers*. For the buffer size of 128 KB, the throughput reaches its peak at 6 *Event Handlers* with a maximum throughput of around 550,000 eps. The results of this phase demonstrated that the optimal configuration parameters for a DP are 64 KB for buffer size and 8 threads of *Event Handlers*. Given the characteristics of the machine used for the experiments, having beyond 6-8 *Event Handlers* causes performance degradation and scheduling overhead, moreover, not all threads would be actively running.

Phase 2 - Throughput Performance Testing using Optimal Parameter Combination.

The goal of this phase was to carry out a first round of performance testing of our architecture using the optimal combination obtained in Phase 1. The data load generated by the client was 70 million event logs containing network details. The IP address was the data pattern defined as sensitive information in the *Rules Descriptor* file and the anonymization technique applied was Base64 encoding. In order to reduce the cost of the output operation, the anonymized output stream was not stored. Table 3 shows the results for the measured metrics: memory usage, CPU usage and throughput.

The results of this phase showed that the HPAE can achieve a high throughput (more than 400,000 eps, reaching a peak of more than 560,000 eps) with low resources (less than 35% CPU usage and a maximum of 102 MB of memory) using a relatively modest test machine.

5 Conclusions and Future Work

Security remains to be the primary obstacle preventing the adoption of cloud services, inside which data protection is one of the top concerns.

In this paper, we presented the HPAE, a practical approach that will enable organizations to implement their own controls to protect sensitive information from service providers in a public cloud environment. We demonstrated that the approach is practical by implementing a prototype and then validated its feasibility to efficiently handle large volumes of data. In our experiments, we achieved a throughput of more than 560,000 eps, using less than 35% of CPU and only 102 MB of memory.

Future work will focus on the integration of the remaining elements of our approach into our prototype, as well as extending the number of available anonymization techniques with a special emphasis on assessing the performance overhead they might introduce to the system. Furthermore, high performance will continue to be a key objective, thus investigating other possible mechanisms to increase throughput and reduce response time will also be part of future work. Other interesting aspects would be to automate the identification of sensitive information. This could be achieved by using self-learning algorithms that discover new data patterns on-the-fly, hence facilitating the automatic configuration of privacy policies.

Acknowledgements

This work was supported, in part, by Science Foundation Ireland grant 10/CE/I1855 to Lero - the Irish Software Engineering Research Centre (www.lero.ie). This work has also received support from Science Foundation Ireland (SFI) via grant 08/SRC/I1403 FAME SRC (Federated, Autonomic Management of End-to-End Communications Services - Scientific Research Cluster).

References

[1] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," 2011. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

[2] Deloitte, "CIO Survey Report 2012 Leading the transformation to a digital future," 2012. [Online]. Available: http://www.deloitte.com/assets/Dcom-Ireland/LocalAssets/Documents/Consulting/IE_Deloitte_CIO_Survey_2012.pdf

[3] J. G. Harris and A. E. Alter, "Cloud Rise: Rewards and Risks at the Dawn of Cloud Computing," 2010. [Online]. Available: http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture_Cloudrise_Rewards_and_Risks_at_the_Dawn_of_Cloud_Computing.pdf

[4] K. Kessinger and M. Gellman, "2010 ISACA IT Risk/Reward BarometerUS Edition," 2010. [Online]. Available: http://www.isaca.org/About-ISACA/Press-room/News-Releases/2010/Documents/2010_ISACA_Risk_Reward_Barometer_Results_US.pdf

[5] F. Gens, "New IDC IT Cloud Services Survey: Top Benefits and Challenges," 2009. [Online]. Available: http://blogs.idc.com/ie/?p=730

[6] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," *Journal of Network and Computer Applications*, vol. 34, pp. 1–11, Jan. 2011.

[7] M. Zhou, R. Zhang, W. Xie, W. Qian, and A. Zhou, "Security and Privacy in Cloud Computing: A Survey," *6th International Conf. on Semantics, Knowledge and Grids*, pp. 105–112, Nov. 2010.

[8] R. Chow, P. Golle, and M. Jakobsson, "Controlling data in the cloud: outsourcing computation without outsourcing control," in *ACM Workshop on Cloud Computing Security*, Chicago, IL, 2009.

[9] iSMG, "Overcoming the Apprehension of Cloud Computing," 2012. [Online]. Available: http://docs.ismgcorp.com/files/handbooks/Cloud-Survey-2012/Cloud_Survey_Report_2012.pdf

[10] Cloud Security Alliance, "Top Threats to Cloud Computing," 2010. [Online]. Available: https://cloudsecurityalliance.org/topthreats/csathreats.v1.0.pdf

[11] Ponemon Institute, "2009 Annual Study: Global Cost of a Data Breach," 2009. [Online]. Available: http://www.securityprivacyandthelaw.com/uploads/file/Ponemon_COB_2009_GL.pdf

[12] S. et al., "Data Breach Trends & Stats," 2012. [Online]. Available: http://www.indefenseofdata.com/data-breach-trends-stats/

[13] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *IEEE Symposium on Security and Privacy*. IEEE, 2000, pp. 44–55.

[14] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," *Advances in Cryptology Eurocrypt*, no. 3027, pp. 506–522, 2004.

[15] C. Gentry, "A fully homomorphic encryption scheme," Ph.D. thesis, Stanford University, 2009.

[16] K. Puttaswamy, C. Kruegel, and B. Zhao, "Silverline: toward data confidentiality in storage-intensive cloud applications," 2011.

[17] M. Mowbray, S. Pearson, and Y. Shen, "Enhancing privacy in cloud computing via policy-based obfuscation," *The Journal of Supercomputing*, pp. 267–291, Mar. 2010.

[18] S. S. Yau and H. G. An, "Protection of users' data confidentiality in cloud computing," *Proceedings of the Second Asia-Pacific Symposium on Internetware*, pp. 1–6, 2010.

[19] LMAX-Exchange, "LMAX Disruptor High Performance Inter-Thread Messaging Library." [Online]. Available: http://lmax-exchange.github.com/disruptor/

[20] L. Dignan, "Software development budgets on the rise, study finds." [Online]. Available: http://www.zdnet.com/software-development-budgets-on-the-rise-study-finds-3040092512/

[21] K. Kent and M. Souppaya, "Guide to Computer Security Log Management," 2006. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-92/SP800-92.pdf

[22] M. Thompson, D. Farley, M. Barker, P. Gee, and A. Stewart, "Disruptor: High performance alternative to bounded queues for exchanging data between concurrent threads," 2011. [Online]. Available: http://disruptor.googlecode.com/files/Disruptor-1.0.pdf

Towards Determining Web Browsing Quality of Experience

Liam Fallon [1], Declan O'Sullivan [2]

[1] Network Management Lab, LM Ericsson, Athlone, Co. Westmeath, Ireland
liam.fallon(at)ericsson.com

[2] Knowledge & Data Engineering Group (KDEG), Trinity College Dublin, Ireland
declan.osullivan(at)cs.tcd.ie

Abstract

Web browsing has emerged as one of the most important end user services carried on today's networks. However, approaches for determining the Quality of Experience of end user sessions are immature, due to the myriad of factors that affect user perception of web browsing. This paper describes initial work towards a framework for determining end user service quality. It describes an approach for measuring the performance of web browsing sessions and presents measurement results obtained using that approach.

Keywords: Quality of Service, Quality of Experience, Web Browsing

1 Introduction

Web browsing is one of the most important services carried on today's networks and has overtaken peer-to-peer file transfer services in traffic use, generating over 50% of Internet traffic [Maier et al., 2009]. The growth in web browsing traffic load is largely driven by the use of HTTP to carry services such as YouTube and TV catch up [Ihm and Pai, 2011].

The TM Forum [TMF, 2009] have pointed out the importance of focusing on managing the Quality of Experience perceived by the consumer of a service rather than on the raw performance metrics of the network delivering that service. For web browsing, that Quality of Experience is affected by many factors [Toutain et al., 2011] such as the reason the user is loading the web site, the expectation that particular user has on delay, and the expectation that user has on the provider of the web site [Bouch et al., 2000], as well as the performance of the underlying network. Determining Quality of Experience in a quantitive manner is a difficult task [Stankiewicz et al., 2011] because perception of quality is subjective and is influenced by many factors.

This paper describes initial work carried out by us towards defining a framework for quantitive determination of web browsing quality of experience, a framework that can be used to assess the QoE of web browsing on access networks. We have developed an approach for measuring web browsing quality that measures web browsing performance at the browser engine, the nearest point possible to the end user. We applied that measurement approach to conduct tests that give indications as to what factors are most important to consider when developing our framework.

The paper is structured as follows. Section 2 presents related work. In Section 3 we describe our approach for measuring web browsing quality. The measurements from tests conducted using that approach are presented in Section 4. Section 5 presents our conclusions and describes future work to be undertaken.

2 Related Work

2.1 Web Browsing Quality

The ITU-T recommends an empirical opinion model for web browsing applications based purely on session time [ITU-T, 2005]. The recommendation describes three perceptual regions for response time as being Instantaneous ($\leq 0.1s$), Uninterrupted ($\leq 1.0s$) and Loss of Attention ($> 10s$). The recommendation observes that, for all session durations *"perceived quality goes down linearly with the logarithm of the session time"*, and recommends the use of the equations in Table 1 for calculating the MOS (Mean Opinion Score) of short, medium, and long web browsing sessions on fast network connections. The ITU-T model must be used with caution because the raw performance of a web page is not the only criterion that end users will use to assess their service experience. However, in cases where different sessions to the same web site are being compared, the equation is applicable.

Page load times, the total time to load all the information on a page, appear more frequently than MOS values when assessing the performance of web browsing services, [Sundaresan et al., 2011] and [OFCOM, 2012] both use page load times of a test web page as a service experience metric for web browsing. Although web browsing perception is affected by many factors [Toutain et al., 2011], objective values for web browsing quality metrics such as those listed in Table 1 can be specified to set a formal expectation on what quality is expected for web browsing.

2.2 How Network Impairments Affect Web Browsing Sessions

The [Sundaresan et al., 2011] study examined the influence of changes in throughput and latency on the download time of small web pages with an average size of 125KB. As one would expect, there is a relationship between throughput and page load time. At throughputs greater than 10 Mbit/s, the page load time remains constant at about 0.8s. As throughput decreases, page load time increases, taking about 1s at a throughput of 5Mbit/s. At 2Mbit/s, page load increases to 1.3s, increasing to 1.7s at 1Mbit/s. The results also show a roughly linear relationship between page load time and latency. Page load time is around 0.8s with a latency of 10ms. It increases to about 1.1s as latency increases to 50ms. Minimum download time is more sensitive to latency, increasing by about 50% from 0.45s to 0.75s as latency increases from 10ms to 40ms.

The [OFCOM, 2012] report shows average web page load times of about 0.7s and peak load times of about 2.2s on connections advertised as being up to 20Mbit/s. The measurements were collected using similar methodology to that used in Sundaresan et al. and are consistent with that study's results.

[Ihm and Pai, 2011] used five years of real data collected from a web proxy system to study web traffic from four countries. They observe a doubling of the average size and average number of objects on web pages between 2006 and 2010. To compensate for this increasing size and complexity, modern web browsers use parallel sessions to optimize page loading. Parallel sessions are used to load elements on a page, with observations from 2010 showing 50% of users holding more than 5 concurrent sessions while browsing the web. In 2010, average page loading latency was 6s, 9s, 10.5s, and 13s for the four studied countries.

[Staehle et al., 2008] studied the influence of packet loss and delay on user tasks such as typing, scrolling, and selecting on office applications running in a thin client using Citrix products. The results

Table 1: Web Browsing Quality of Experience Parameters

Parameter	**Values**	**Source**
MOS ($sessiontime < 10s$)	$4.38 - 1.30.ln(sessiontime)$	[ITU-T, 2005]
MOS ($10s \leq sessiontime < 30s$)	$4.79 - 1.03.ln(sessiontime)$	[ITU-T, 2005]
MOS ($30s \leq sessiontime$)	$5.76 - 0.948.ln(sessiontime)$	[ITU-T, 2005]
Page Load Time	Seconds	[Sundaresan et al., 2011] [OFCOM, 2012]

show that there is a linear relationship between both delay and packet loss and the amount of time it takes to complete a task on a thin client, with delay generally having a bigger influence than packet loss. One finding was that a delay increase of 200ms could increase the time taken to execute a menu action by 30%. A user study carried out by the authors showed that an increase in time taken to execute tasks correlated with a decrease in user satisfaction *"to a satisfying degree"*.

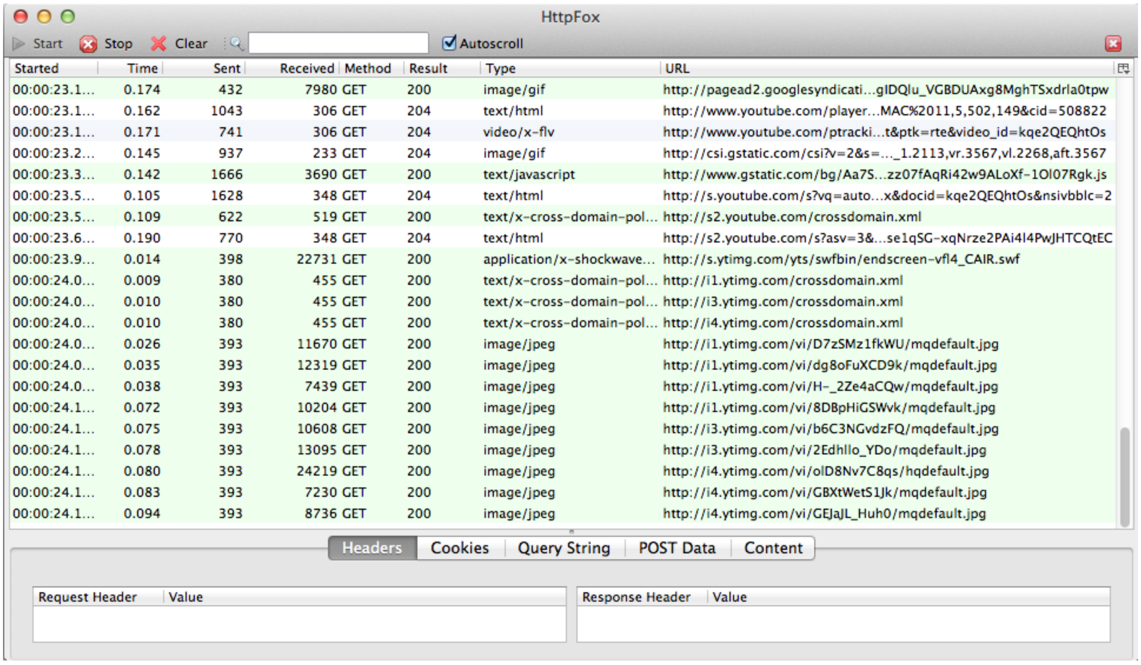
3 An Approach for Measuring Web Browsing Quality

In order to determine how web browsing sessions behave and to determine what contextual factors are the most important for web browsing, one must have a mechanism for quantitively measuring the performance of those web browsing sessions. The widely used *wget* [GNU, 2013] downloads pages on a single HTTP request-response basis, and *Httpref* [Hewlitt Packard, 2013] is built to benchmark the performance of web servers. Because web browsers run many requests from a single page in parallel and also heavily employ caching to improve performance, the most practical approach to determine the quality of the web browsing experience is to measure at the web browser engine. We selected the *Firefox* web browser [Mozilla, 2013] for measurement because it is the most commonly used open source browser and it has a rich ecosystem of open source plugins that are available to adapt.

The *HttpFox* Firefox plugin [Martin Theimer, 2013], an open source application written in JavaScript, records, analyses and presents the performance of a web page load graphically, showing the parallelism in the loading and also the degree of caching used. It gives an interactive picture of the user-perceived browsing experience. The user interface of the plugin is shown in Figure 1. The HttpFox plugin shows results only on its user interface, so we enhanced it to write its page loading analysis to file.

The *iMacros* Firefox plugin [iOpus, 2013] allows Firefox to be scripted and to run under the control of a macro. A sequence of web pages can be specified for loading, delays can also be introduced between web page loads, and loops can be specified. Plugins like iMacros are commonly used for web scraping. In Figure 2, a Firefox session running a YouTube video under the control of iMacros is shown.

In our measurement approach shown in Figure 3, a macro for iMacros is written that specifies the web pages to load. Firefox is executed under the control of that macro, and HttpFox saves web session metrics to file. That log file contains page load time measurements, the volume of data downloaded and



Started	Time	Sent	Received	Method	Result	Type	URL
00:00:23.1...	0.174	432	7980	GET	200	image/gif	http://pagead2.googlesyndicati...gIDQlu_VGBDUAxg8MghTSxdrla0tpw
00:00:23.1...	0.162	1043	306	GET	204	text/html	http://www.youtube.com/player...MAC%2011,5,502,149&cid=508822
00:00:23.1...	0.171	741	306	GET	204	video/x-flv	http://www.youtube.com/ptracki...t&ptk=rte&video_id=kqe2QEQhtOs
00:00:23.2...	0.145	937	233	GET	204	image/gif	http://csi.gstatic.com/csi?v=2&s=..._1.2113,vr.3567,vl.2268,aft.3567
00:00:23.3...	0.142	1666	3690	GET	200	text/javascript	http://www.gstatic.com/bg/Aa7S...zz07fAqRi42w9ALoXf-1Ol07Rgk.js
00:00:23.5...	0.105	1628	348	GET	204	text/html	http://s.youtube.com/s?vq=auto...x&docid=kqe2QEQhtOs&nsivbblc=2
00:00:23.5...	0.109	622	519	GET	200	text/x-cross-domain-pol...	http://s2.youtube.com/crossdomain.xml
00:00:23.6...	0.190	770	348	GET	204	text/html	http://s2.youtube.com/s?asv=3&...se1qSG-xqNrze2PAi4l4PwJHTCQtEC
00:00:23.9...	0.014	398	22731	GET	200	application/x-shockwave...	http://s.ytimg.com/yts/swfbin/endscreen-vfl4_CAIR.swf
00:00:24.0...	0.009	380	455	GET	200	text/x-cross-domain-pol...	http://i1.ytimg.com/crossdomain.xml
00:00:24.0...	0.010	380	455	GET	200	text/x-cross-domain-pol...	http://i3.ytimg.com/crossdomain.xml
00:00:24.0...	0.010	380	455	GET	200	text/x-cross-domain-pol...	http://i4.ytimg.com/crossdomain.xml
00:00:24.0...	0.026	393	11670	GET	200	image/jpeg	http://i1.ytimg.com/vi/D7zSMz1fkWU/mqdefault.jpg
00:00:24.0...	0.035	393	12319	GET	200	image/jpeg	http://i1.ytimg.com/vi/dg8oFuXCD9k/mqdefault.jpg
00:00:24.0...	0.038	393	7439	GET	200	image/jpeg	http://i1.ytimg.com/vi/H-_2Ze4aCQw/mqdefault.jpg
00:00:24.1...	0.072	393	10204	GET	200	image/jpeg	http://i1.ytimg.com/vi/8DBpHiGSWvk/mqdefault.jpg
00:00:24.1...	0.075	393	10608	GET	200	image/jpeg	http://i3.ytimg.com/vi/b6C3NGvdzFQ/mqdefault.jpg
00:00:24.1...	0.078	393	13095	GET	200	image/jpeg	http://i3.ytimg.com/vi/2Edhllo_YDo/mqdefault.jpg
00:00:24.1...	0.080	393	24219	GET	200	image/jpeg	http://i4.ytimg.com/vi/olD8Nv7C8qs/hqdefault.jpg
00:00:24.1...	0.083	393	7230	GET	200	image/jpeg	http://i4.ytimg.com/vi/GBXtWetS1Jk/mqdefault.jpg
00:00:24.1...	0.094	393	8736	GET	200	image/jpeg	http://i4.ytimg.com/vi/GEJaJL_Huh0/mqdefault.jpg

Figure 1: The HttpFox Firefox Plugin

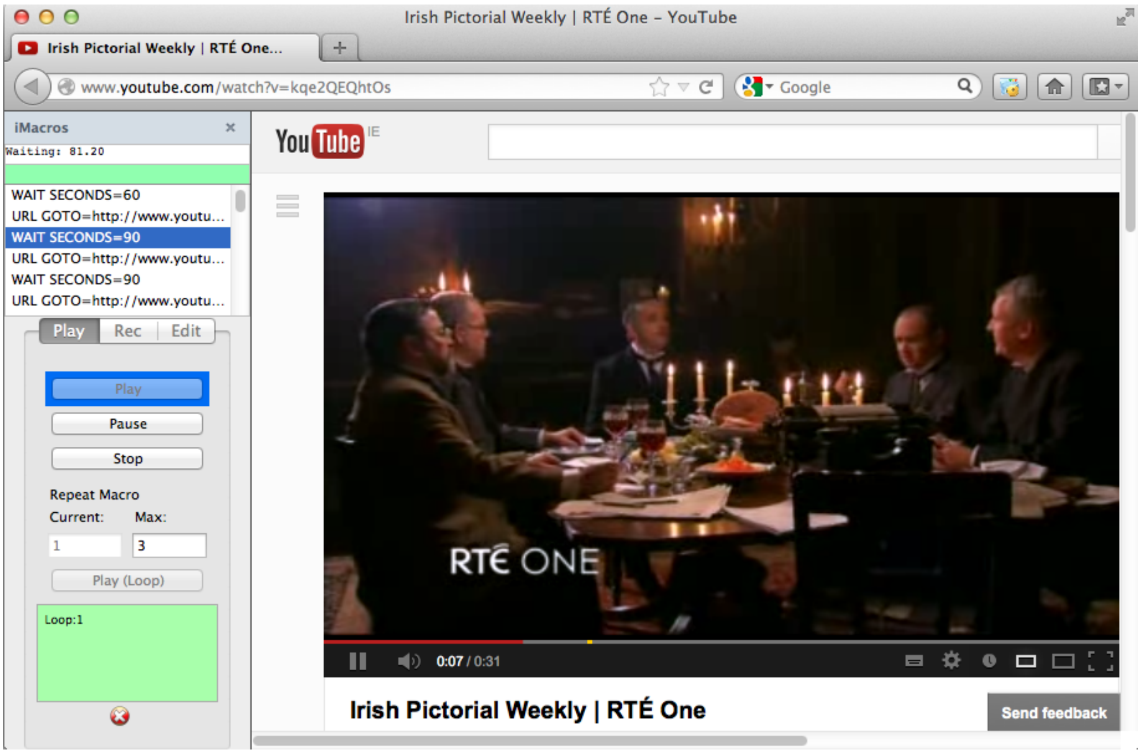

Irish Pictorial Weekly | RTÉ One - YouTube
Irish Pictorial Weekly | RTÉ One...
www.youtube.com/watch?v=kqe2QEQhtOs
Google
iMacros
Waiting: 81.20
WAIT SECONDS=60
URL GOTO=http://www.youtu...
WAIT SECONDS=90
URL GOTO=http://www.youtu...
WAIT SECONDS=90
URL GOTO=http://www.youtu...
Play
Rec
Edit
Play
Pause
Stop
Repeat Macro
Current:
Max:
1
3
Play (Loop)
Loop:1
RTÉ ONE
0:07 / 0:31
Irish Pictorial Weekly | RTÉ One
Send feedback

Figure 2: Firefox running under control of the iMacros Plugin

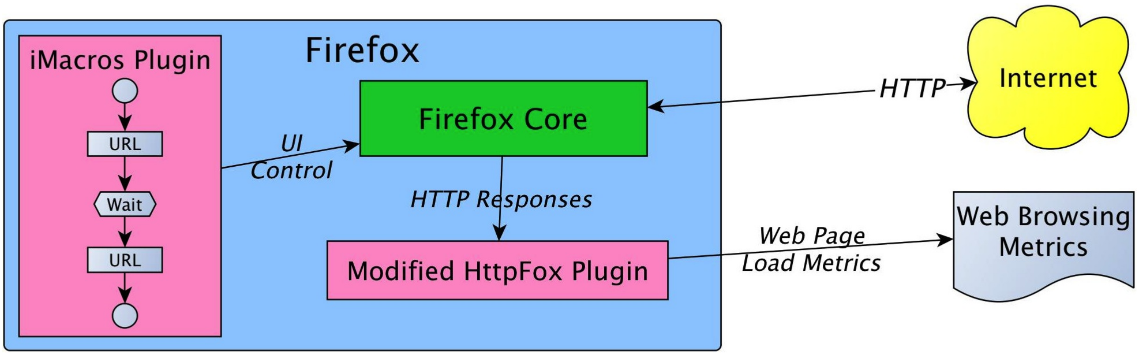


Figure 3: Capturing Web Browsing Experience

uploaded, details on each HTTP request and the status of each completed request.

4 Web Browsing Quality Measurements

We carried out tests to obtain preliminary measurements of page load time on a selected a set of web sites on various access types. Firefox caching was disabled in order to force all web page requests to load from the network. Firefox was configured to execute 32 HTTP requests in parallel, the default value.

The characteristic of each web site selected for measurement are shown in Table 2. The web sites were chosen because they range from small to large in size and represent a mix of the types of web sites commonly used. The *Page Loads* column gives the number of page loads considered in calculating average and standard deviation values of the other metrics over all tests on all access types. The *KB Uploaded* and *KB Downloaded* columns give the average and standard deviation of the amount of data uploaded to and downloaded from each web site. The *Request Count* column shows the average and standard deviation of the amount of HTTP requests used to load each page. The characteristics show that *www.google.com* is the least complex web page, and has the smallest download size. Unsurprisingly,

Table 2: Web Site Characteristics

URL	Type	Page Loads	KB Uploaded		KB Downloaded		Request Count	
			Avg	Stdev	Avg	Stdev	Avg	Stdev
www.ait.ie	Education	49	26.1	1.2	801.6	646.6	46.4	2.5
www.ericsson.com	Corporate	56	26.8	3.3	970.5	1054.8	45.6	5.8
www.google.com	Search Engine	59	9.7	1.8	353.1	111.0	16.5	3.1
www.irishtimes.com	Newspaper	32	137.2	3.8	1909.0	116.7	214.2	3.4
www.linkedin.com	Social Network	41	71.9	14.7	709.9	267.6	120.3	20.8
www.youtube.com	Streaming Video (31 Second Clip)	32	66.4	4.5	3078.9	302.3	97.8	4.7

www.youtube.com has the largest download size. The most complex web page, the page with the highest number of HTTP requests is *www.irishtimes.com*.

Table 3 shows access types over which measurements were conducted. The network metrics in the table are the results of speed tests conducted just prior to our tests being run. The tests on the *Fixed Wireless* access type were run at a time of peak load and at a time of low load. The tests on *Mobile Broadband* were conducted at a location with a strong signal during the afternoon and at a location with a marginal signal at midnight; time of day is likely to have led to higher download and upload speeds observed with a weak signal. Tests for *www.youtube.com* and *www.irishtimes.com* were not conducted on the *Rural DSL* access type because the load time of those sites was many minutes on that access type.

The chart in Figure 4 shows the load times for each web site in Table 3 for each access type in Table 3. The full results appear in Table 4 in Appendix I.

As one would expect, the *Corporate LAN* access type performed best on all web sites. However, even though this access type is more than 10 times faster than the next fastest access type, page load speeds on this access type are not 10 times faster. This is because other factors such as page serving time on web sites, the use of web proxies, and delays in the Internet contribute to page load time.

The *Rural DSL* access type had the worst performance; this connection must be considered to be a marginal connection even for the most simple web sites and is not discussed further in this section.

The four access types performed adequately on the web sites *www.google.com* and *www.ait.ie*, showing page load times of one to five seconds. It is interesting to note the difference in load times for *www.ait.ie* and *www.ericsson .com*. Those two web sites have almost identical characteristics, but *www. ericsson.com*, which is served from Sweden, shows longer page load times of eight to fifteen seconds on Irish access networks, indicating that access type is but one network factor affecting page load times.

The results indicate that the number of HTTP requests required for a web page load is an important factor that affects load time. The *www.linkedin.com* web site is smaller than both *www.ait.ie* and *www.ericsson .com* but takes between 21 and 52 seconds to load. A load of the web page executes 120 HTTP requests, three times more than *www.ait.ie* and *www.ericsson.com*. Firefox was configured to execute 32 HTTP requests in parallel, so web pages with more requests will have longer load times.

The mobile broadband test results show better performance than the equivalent fixed wireless tests. The mobile broadband performance over a weak signal are only marginally worse than performance over

Table 3: Access Network Characteristics (www.speedtest.net)

Access Type	Ping Time	Download Speed	Upload Speed
Corporate LAN	10ms	50.54Mbit/s	56.38Mbit/s
Fixed Wireless (Busy Peak)	72ms	2.94Mbit/s	0.05Mbit/s
Fixed Wireless (Off Peak)	75ms	3.46Mbit/s	0.37Mbit/s
Mobile Broadband (Strong Signal)	76ms	2.41Mbit/s	0.57Mbit/s
Mobile Broadband (Weak Signal)	83ms	3.63Mbit/s	2.41Mbit/s
Rural DSL	85ms	0.13Mbit/s	0.21Mbit/s

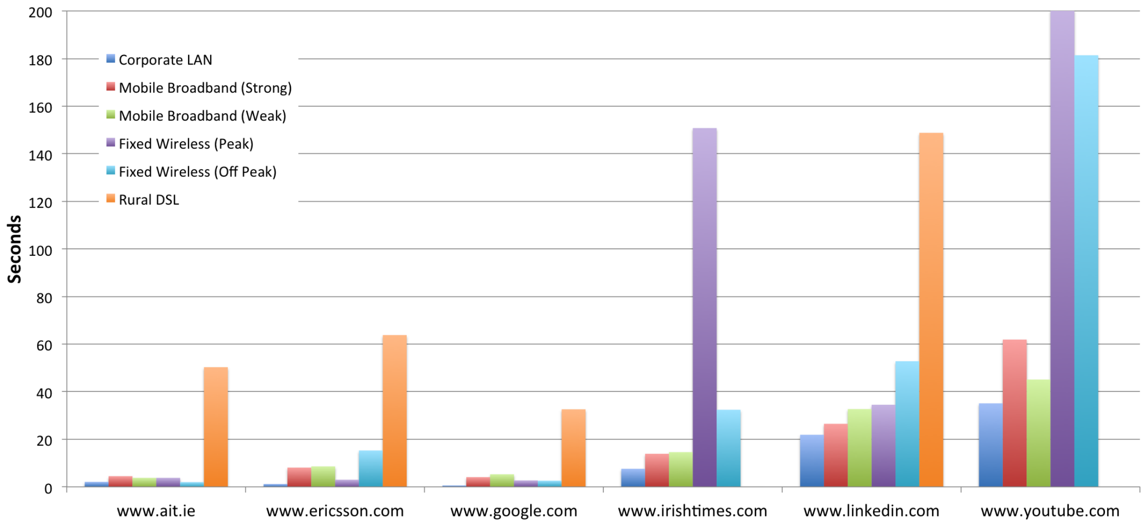


Figure 4: Page Load Times for Web Sites by Access Type

a strong signal for most web sites, and for *www.youtube.com* they are marginally better.

The fixed wireless test results show poor performance in comparison with mobile broadband on complex and large web sites. The abort and error counts (see Table 4 in Appendix I) indicate that the browser was encountering problems in loading these web sites over fixed wireless access. The upload speed on the *Fixed Wireless (Peak)* was very low during tests on that access type, and may be a factor that affects those load times.

5 Conclusions and Future Work

This paper describes our initial work towards a framework that can be used to assess the web browsing service quality of an access network. It explains our approach for measuring web browsing session performance, and presents results of tests carried out using that approach.

In order to develop a framework to determine web browsing quality, the following challenges remain.

1. The type of web site affects end user perception. Although load time is an important measurement, other factors such as the type of web site, and the order of request download must also be taken into account. The download times of a streaming web site is fast enough if it can keep up with video play out rate, and users will not complain if only requests that download adverts are slow.

2. The structure of the web site must be considered. The web page size and number of requests to load a page are just two factors considered in this paper. Other structural factors such as pages using Flash or highly interactive pages such as *Google Docs* have not been considered.

3. Benchmark web sites that cover the main types of services carried over HTTP are needed. Commercial web sites are regularly updated in an uncontrolled manner and cannot be used as benchmarks. Types of benchmark web sites might be simple, complex, newspaper, social networking, TV catch up, or video streaming.

4. The type of device on which pages are being rendered must be considered because that alters the users perception of quality.

Acknowledgment

This work was partially supported by Science Foundation Ireland under the FAME SRC Grant No 08/SRC/I1403.

Appendix I

Table 4: Page Load Times by Access Type

URL	Access	Page Load Time		Abort	Error
	Type	Avg	Stdev	Count	Count
www.ait.ie	Corporate LAN	2.1	4	0	4
www.ait.ie	Fixed Wireless (Off Peak)	2	0.3	0	4
www.ait.ie	Fixed Wireless (Peak)	3.8	4.2	0	4
www.ait.ie	Mobile Broadband (Strong)	4.5	1.5	0	0
www.ait.ie	Mobile Broadband (Weak)	3.8	1	0	4
www.ait.ie	Rural DSL	50.3	6.6	0	4
www.ericsson.com	Corporate LAN	1.2	0.2	0	0
www.ericsson.com	Fixed Wireless (Off Peak)	15.3	40.1	0	0
www.ericsson.com	Fixed Wireless (Peak)	3	0.9	0	0
www.ericsson.com	Mobile Broadband (Strong)	8.1	7.5	0	0
www.ericsson.com	Mobile Broadband (Weak)	8.6	8.5	0	0
www.ericsson.com	Rural DSL	63.8	11.2	0	0
www.google.com	Corporate LAN	0.6	0.1	0	0
www.google.com	Fixed Wireless (Off Peak)	2.6	3.9	0	0
www.google.com	Fixed Wireless (Peak)	2.7	4.3	0	0
www.google.com	Mobile Broadband (Strong)	4.1	2.9	0	0
www.google.com	Mobile Broadband (Weak)	5.3	6.1	0	0
www.google.com	Rural DSL	32.6	9.4	0	0
www.irishtimes.com	Corporate LAN	7.6	1.6	0	0
www.irishtimes.com	Fixed Wireless (Off Peak)	32.4	53	1	21
www.irishtimes.com	Fixed Wireless (Peak)	150.8	238.8	6	5
www.irishtimes.com	Mobile Broadband (Strong)	13.9	1.9	0	16
www.irishtimes.com	Mobile Broadband (Weak)	14.6	7.4	0	2
www.irishtimes.com	Rural DSL	0		0	0
www.linkedin.com	Corporate LAN	21.9	10.6	0	0
www.linkedin.com	Fixed Wireless (Off Peak)	52.8	84.5	5	24
www.linkedin.com	Fixed Wireless (Peak)	34.5	40.1	0	0
www.linkedin.com	Mobile Broadband (Strong)	26.5	10.7	0	0
www.linkedin.com	Mobile Broadband (Weak)	32.7	23.2	0	0
www.linkedin.com	Rural DSL	148.8	82.7	4	0
www.youtube.com	Corporate LAN	35.1	3.9	0	0
www.youtube.com	Fixed Wireless (Off Peak)	181.4	230.2	263	4
www.youtube.com	Fixed Wireless (Peak)	200.5	269.5	241	9
www.youtube.com	Mobile Broadband (Strong)	61.9	43.1	12	0
www.youtube.com	Mobile Broadband (Weak)	45.1	6.4	1	0
www.youtube.com	Rural DSL	0		0	0

References

[Bouch et al., 2000] Bouch, A., Kuchinsky, A., and Bhatti, N. (2000). Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '00, pages 297–304, New York, NY, USA. ACM.

[GNU, 2013] GNU (2013). Gnu wget.

[Hewlitt Packard, 2013] Hewlitt Packard (2013). Httperf tool for measuring web server performance.

[Ihm and Pai, 2011] Ihm, S. and Pai, V. S. (2011). Towards understanding modern web traffic. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, IMC '11, pages 295–312, New York, NY, USA. ACM.

[iOpus, 2013] iOpus (2013). A firefox automator.

[ITU-T, 2005] ITU-T (2005). Estimating end-to-end performance in ip networks for data applications. Technical Report G.1030, ITU-T.

[Maier et al., 2009] Maier, G., Feldmann, A., Paxson, V., and Allman, M. (2009). On dominant characteristics of residential broadband internet traffic. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 90–102, New York, NY, USA. ACM.

[Martin Theimer, 2013] Martin Theimer (2013). An http analyzer addon for firefox.

[Mozilla, 2013] Mozilla (2013). The firefox web browser.

[OFCOM, 2012] OFCOM (2012). Uk fixed-line broadband performance, nov 2011. Report, Office of Communications.

[Staehle et al., 2008] Staehle, B., Binzenhoefer, A., Schlosser, D., and Boder, B. (2008). Quantifying the influence of network conditions on the service quality experienced by a thin client user. *Measuring, Modelling and Evaluation of Computer and Communication Systems (MMB), 2008 14th GI/ITG Conference* -, pages 1 –15.

[Stankiewicz et al., 2011] Stankiewicz, R., Cholda, P., and Jajszczyk, A. (2011). Qox: What is it really? *IEEE Communications Magazine*, 49(4):148 –158.

[Sundaresan et al., 2011] Sundaresan, S., de Donato, W., Feamster, N., Teixeira, R., Crawford, S., and Pescapè, A. (2011). Broadband internet performance: A view from the gateway. *SIGCOMM Comput. Commun. Rev.*, 41(4):134–145.

[TMF, 2009] TMF (2009). Technical report: Managing the quality of customer experience (mce). Technical Report TR148, TM-Forum.

[Toutain et al., 2011] Toutain, F., Bouabdallah, A., Zemek, R., and Daloz, C. (2011). Interpersonal context-aware communication services. *IEEE Communications Magazine*, 49(1):68 –74.

Session 5

Networking II

Building a Scalable Event Processing System with Messaging and Policies – Test and Evaluation of RabbitMQ and Drools Expert

Sumit Dawar[1, 2], Sven van der Meer[1], Enda Fallon[2], John Keeney[1], Tom Bennett[2]

[1]Ericsson Network Management Labs, Ericsson Software Campus, Athlone, Co. Westmeath
<sumit.dawar@ericsson.com>

[2]Athlone Institute of Technology, Athlone, Co. Westmeath

Abstract

This paper describes an architecture, implementation and performance tests for a policy-based event processing system. The main advantage of our approach is that we use policies for event pattern matching (an advanced form of Complex Event Processing) and for the selection of corrective actions (called Distributed Governance). Policies are (a) distributed (over multiple components) and (b) coordinated (using centralized authoring). The resulting system can deal with large numbers of incoming events, as is required in a telecommunication environment. Peak load will be well above 1 million events per second, combining different data sources of a mobile network. This paper presents the motivation for such a system, along with a comprehensive presentation of its design, implementation and evaluation.

Keywords: Advanced Message Queuing Protocol (AMQP), Rules Engine, Event Processing.

1 Introduction

Mobile networks are growing, cell sizes are decreasing and the number of connected devices is exploding. These conditions result in an ever increasing number of events from the network. The situation becomes critical and requires scalable solutions for event processing and the selection of corrective actions, i.e. for alarm events. Our work combines rule systems to encode event processing knowledge and messaging to provide for a distributed system. Other earlier work used centralized rules over a distributed system [5-8], which drastically limits the scalability. We use distributed rules that are coordinated by centralized authoring to address this limitation. In this paper, we describe the general architecture, the reference implementation we have developed and performance tests with regard to end-to-end message processing. The work is integrated into a wider research project in the Ericsson Network Management labs that deals with hundreds of thousands of events per second.

This paper is organized as follows: section 2 introduces core concepts, technologies and products from messaging systems and rule systems. Section 3 discusses related work from academia and industry. Section 4 briefly discusses the architecture and main design decisions of our system. The sections 5 and 6 then detail the implementation and provide a discussion of test results, mainly looking into the performance for the end-to-end event processing. Finally, a conclusion summaries this paper and discusses future work items of our project.

2 Conceptual Background, Products and Tools

Combining concepts from messaging systems with concepts from rule systems requires an understanding of two disjoint domains. In general, messaging system provides the main communication links between the components of a distributed system. A rule system provides the intelligence to manage and process events and event patterns to trigger appropriate actions. In this section we look into the fundamental idea of both to introduce relevant terms and concepts.

2.1 Messaging System

A distributed system has multiple components that may be built independently, with potentially different languages and platforms, dispersed at different locations. There are a number of approaches including: distributed data stores, streamed data, query-response models, or asynchronous messaging. Using a message-based approach distributed components share and process data in a responsive asynchronous way and it is this approach we focus on in this work. Our works use *Advanced Message Queuing Protocol* (AMQP) messaging due to external project requirements, namely RabbitMQ, an open source AMQP implementation.

AMQP [1] is an open standard for passing business messages between applications. Data (the messages) is sent in a stream of octets, thus it is often called a 'wire protocol'. Version 1.0 of the AMQP standard defines three main components: the networking protocol, a message representation and the semantics of broker services. All of these components address core features such as queuing, routing, reliability and security. Message encoding is separated into links, sessions, channels and connections, with links being the highest level and connections the lowest level of abstraction. A link connects network nodes, also known as distributed nodes in AMQP.

RabbitMQ [2] is an open source implementation of the AMQP standard. It facilitates 'producers' to send messages to 'brokers', which in turn deliver them to 'consumers'. Messages can also be routed, buffered and made persistent, depending on runtime configuration.

AMQP is designed to be programmable, allowing application to configure 'entities' and 'routing schemas'. The three important entities in RabbitMQ realizing the programmability are 'exchange', 'queue' and 'binding'. An exchange receives events from a producer and realizes different routing schemes. A queue is bound to an exchange and handles consumer-specific message reception. A binding defines the rules for message transfer between an exchange and a queue. See [3] for details.

2.2 Rule System

Rule systems provide the means to define and process rules. In our work, we are focusing on Production Rule Systems (PRS) due to external project requirements. The computational model of PRS implements the notion of a set of rules, where each rule has a sensory precondition ("left-hand-side", LHS, or "WHEN" clause) and a consequential action ("right-hand-side", RHS, or "THEN" clause). Rules are also referred to as *productions* and they are the primary form of *knowledge* representation. The rule engine also maintains knowledge-base of facts. When the facts stored satisfy the precondition of a rule, the rule "fires", thus invoking the action part of the rule. Often, the action part of the rule can change the fact knowledge-base, potentially triggering more rules.

Drools Expert is an open source implementation of a PRS. In Drools Expert, *Rules and facts* of a PRS constitute a *knowledge base*. Rules are present in the *production memory* and the facts are kept in a database called *working memory*, which maintains current system *knowledge*. There is an *Inference Engine* based on Charles Forgy's *Rete Algorithm*, which efficiently matches the facts from working memory to conditions of the rules in the production memory.

Also, a conflict resolution is required when there are multiple rules on the *agenda*. As firing a rule may have side effects on working memory, the rule engine needs to know in what order the rules should fire (for instance, firing 'ruleA' may cause 'ruleB' to be removed from the agenda). The default conflict resolution strategies employed by *Drools Expert* are: Salience and LIFO (last in, first out). [4]

3 Related Work

In the *Policy-Based Information Sharing in Publish/Subscribe Middleware* [5] author describes a control of sensitive information system in health care environment. The criticality of information sharing and data access is controlled by rules, precisely hook rules (Postgres SQL). Information that travels on the messaging system is tailored for a particular subscriber, on need-to-know basis. We have found that this paper has similar architecture as of our system with a slightly different implementation. Our system analyzes the patterns inside the incoming messages and modifies the forwarded message to correspond to the identified pattern, whereas it analyses the incoming messages and modifies it for particular subscriber according to information relevant to that subscriber.

A rule-based middleware for business process execution [6] implements rules over messaging middleware to provide a simple and efficient way of describing executable business processes. The complex conditional workflows and enterprise integration patterns are implemented in terms of rules. The Prova rule language and the Rule Markup Language (RuleML) are used to implement rules over an Enterprise Service Bus (ESB).

Policy-driven middleware for self-adaptation of web services compositions [7] focuses on specifying and enforcing monitoring-policies to help in fault detection and corrective adaptation of web services compositions. Since monitoring and corrective action selection is combined in a single policy, this work does not scale well when the number of faults increases drastically. It also does not allow for smart filtering of fault events, which is essential to address high-priority events immediately and add lower-priority events to maintenance reports.

Message oriented middleware with integrated rules engine [8] is a patented invention addressing deficiencies in respect to the management of message oriented middleware. It describes the integration of a rule engine with message-oriented middleware. Their method includes creating a shared memory in the memory of a computer and adding or deleting tokens in the shared memory corresponding to objects such as messages and message queues, created in and removed from, respectively, in a messaging component of message oriented middleware, or topics or subscriptions or log file space for messages queues in the messaging component. The method additionally includes applying rules in a rules engine to the tokens in the shared memory.

Our work differs from the above in that we use distributed and coordinated policies (between two components for event processing and governance), while policy instances in each component are atomic, i.e. do not effect each other. This results in a system that is hugely scalable, since only a combination of event processing policy and governance policy depend on each other.

4 Architecture and Design

We receive events from streams (using other Ericsson software), process them and forward them via queues. Each component employs a rule engine to process events. A typical process is to receive an event or a number of events (pattern) and create/send composite events. The events we process are actual mobile network events, such as performance events (counters) or alarm events. However, for simplification we refer to events as characters, e.g. ‘A’, ‘B’ and ‘C’. Figure 4.1 shows how an incoming event stream (ABABCA…) is directed to a dedicated queue (CEP) and processed.

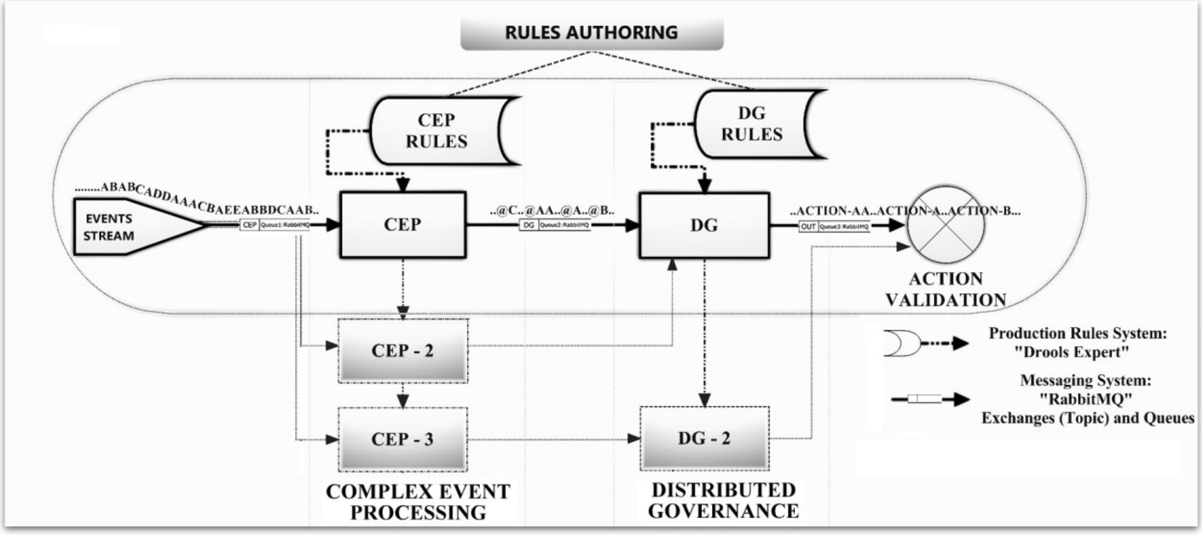


Figure 4.1 Architecture and Deployment Scenarios

Events are received, one by one, by the Complex Event Processing (CEP) component. It takes simple events (‘A’, ‘B’) and generates complex events (‘@A’, ‘@AA’). These complex events represent patterns, i.e. sequences of events that are of special interest. The rules in the CEP component specify which patterns need to be matched and which corresponding complex event needs to be generated. Finally, complex events are sent to the next queue.

The Distributed Governance (DG) component receives complex events and selects appropriate actions to respond to them. The rules in the DG component define which complex events are being processed and what actions are associated with them. The number of associated actions can be zero or more, with zero action indicating an un-decidable situation, while more than one indicates multiple possible actions. DG then sends the actions to a new queue, which can feed into multiple applications of a broader management process, e.g. as part of Network Operation Center (NOC).

Combining messaging (AMQP) and rule systems (PRS) allows for a design of a flexible and scalable system. Using queues for communication not only facilitates the CEP and DG components to be distributed, but also for multiple redundant or load-balanced instances of each component to be run in parallel at runtime. If one CEP instance reaches its performance limits a new CEP instance can be executed, connected to the CEP queue and some patterns of the original CEP instance allocated to the new CEP instance. Figure 4.1 shows a scenario with three CEP instances and two DG instances.

One characteristic of the described system design requires special attention: the processing of patterns and the selection of actions is (a) distributed over two components (CEP and DG in the architecture) and can also be (b) distributed over multiple instances (CEP and DG instances in design and runtime). An effective and efficient coordination is required to guarantee that all patterns are processed and that the resulting complex events find related rules for action selection. Figure 4.1 shows a process for 'Rule Authoring' which is responsible for the coordination. The details of this process are out of scope for this paper, which focuses on the implementation and testing of the message processing.

5 Implementation

This section details the implemented system. We have built four components (which we call nodes), developed in Java 7. Two nodes realize the core of the event processing and two are used to automate tests. The two core nodes are CEP and DG (Figure 5.1). The other two supporting nodes are the input and output consoles (Figure 5.2). CEP and DG are built in a very similar way: they read events (messages) from a topic, invoke a rule engine to process events and then publish the results of the rule evaluation on another topic in form of complex events (CEP) or actions (DG).

5.1 Core Nodes

Both nodes, CEP and DG, start with an initialization of their respective topics and knowledge base (rules, for rule processing). CEP waits to get events from the input console, processes it (applies rules) and sends it out on another topic where DG receives it. Similarly, DG dispatches events with the associated action after processing the received composite event from CEP. This cycle of *waiting* and *processing* goes on endlessly for the core nodes.

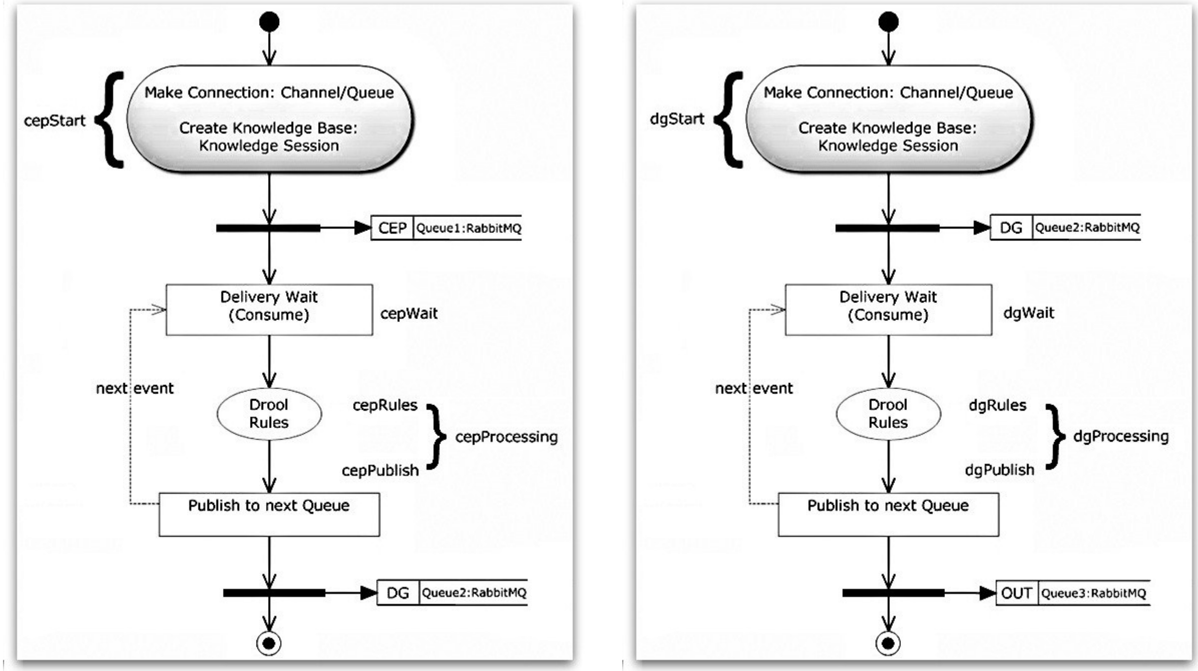


Figure 5.1: Core Nodes, CEP (left) and DG (right)

5.1.1 Complex Event Processing (CEP) Node

Figure 5.1 (left) shows the CEP node with its three main parts: *start*, *wait* and *processing*. Start creates the knowledge base and two topics *CEP* and *DG*. When an event is received on *CEP* topic, a corresponding *fact* is inserted into the knowledge base and all rules are 'fired' (processed). Rules evaluate to match patterns as the knowledge base holds the information (facts) of previously received events. To keep the knowledge base light and efficient these facts are retracted when they are of no use to match patterns. In our system we have kept up to four facts in knowledge base to match the pattern, we call it the window of events. This window size can be changed per event pattern required to be matched. After rules evaluation *complex events* are generated and published to the *DG* topic.

5.1.2 Distributed Governance (DG) Node

Figure 5.1 (right) shows the DG node with its three main parts: *start*, *wait* and *processing*. Similar to the CEP node, DG creates its knowledge base and two topics called *DG* and *OUT*. The topic DG is the same as that created by the CEP node for its output, thus the two nodes a bound via that topic. When a *complex event* is received, a corresponding *fact* is inserted into the knowledge base and all appropriate triggered rules are then fired. Rules evaluate in DG to associate identified patterns to *actions*, which are then published to *OUT* topic.

Table 5.1 shows an example of events (single and multi) and corresponding *complex events* with *associative action*. This pattern matching can be extended to generate new complex events, by simply writing the new CEP rules and corresponding rules in the DG for associative action.

Single event Pattern		
Incoming Event	Composite Event	Associative Action
B	*@B*	*Action-B*
A	*@A*	*Action-A*
C	*@C*	*Action-C*
D	*@D*	*Action-D*
E	*@E*	*Action-E*

Multi-event Pattern		
Incoming Events	Composite Event	Associative Action
A-A	*@AA*	*Action- AA*
A-B†	*@AB*	*Action- AB*
A-A-B	*@AAB*	*Action- AAB*
A-A-B-B	*@AABB*	*Action- AABB*

†*A-B* implies that 'B' occurs after 'A'

Table 5.1: Single and Multi-event Pattern (examples)

5.2 Supporting Nodes

For testing, we have added an input and an output console, which will later be replaced by real systems for *event processing* and *action respond*. For the current system the input node provides the functionality of reading a file containing *events*, and then splitting the string to publish events on the *CEP* topic one by one. The output node receives the *actions* on the *OUT* topic and prints them out.

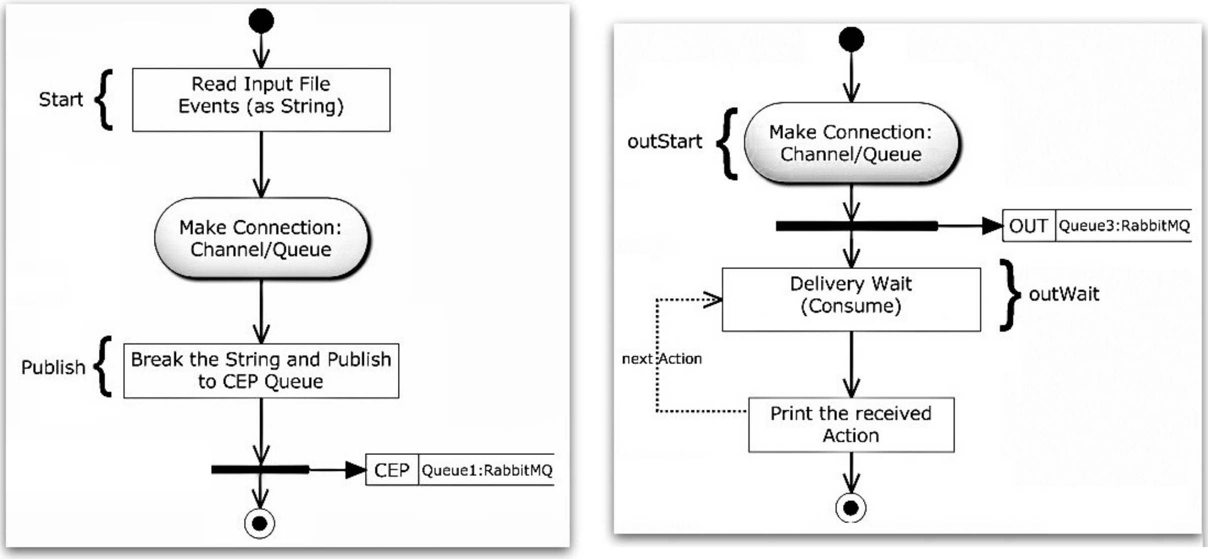

Start
Read Input File Events (as String)
Make Connection: Channel/Queue
Publish
Break the String and Publish to CEP Queue
CEP
Queue1:RabbitMQ
outStart
Make Connection: Channel/Queue
OUT
Queue3:RabbitMQ
Delivery Wait (Consume)
outWait
next Action
Print the received Action

Figure 5.2: Input (left) and Output (right) Console support Nodes

Figure 5.2 shows the two supporting nodes and their main phases (input console on the left and output console on the right). The input console *start*s and *publish*es as described above. When all events read from a file are published, it terminates. The output console *start*s once and *wait*s indefinitely (until the process is terminated). *Start* creates the topic OUT and *wait* waits for actions from the DG node to print them to the console as they arrive.

6 Testing and Evaluation

The tests we have performed are modeled to provide a good understanding about the performance of the overall system. Special attention focuses on the impact the message processing and the rule processing have on the overall system performance. The goal is to understand the technology impact on an end-to-end event processing. Tests have been run for 10 upto 1,000,000 events in a single stream with 10 test runs per input stream size. The number of rules and the actual rules have not been changed between test runs, so the results show the processing of a fixed set of 10 rules for CEP and 9 rules for DG. Further test runs will be needed to understand the impact of increasing rule sets on the performance. All tests have been run on a Intel i5 (dual core) Windows 7 laptop.

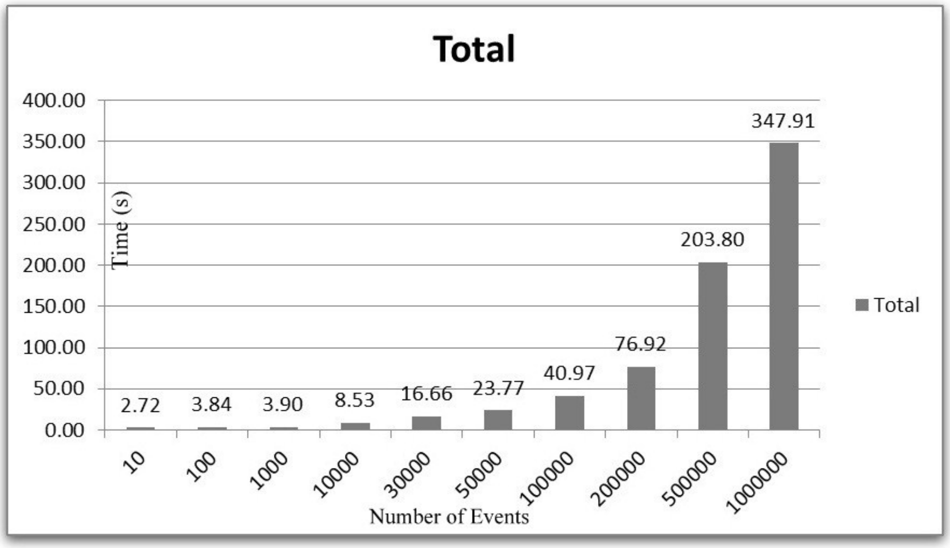


Figure 6.1: Averaged Maximum Time for system

Each component of the system has fixed measurement points. They are shown in the figures in the implementation section. Initialization phases (called *start*) are not part of the measurement. The following list shows all measurement points of each component:

- Core nodes (Figure 5.1): *Start*, *Wait*, *Rules*, *Publish* (for CEP and DG)
- Supporting Nodes (Figure 5.2: *Publish* (Input node) and *Wait* (Output node)

Figure 5.1 shows the overall processing time, i.e. the time it takes to process all events from input console to output console. The time for up to 1,000 events is negligibly small. From 10,000 events onwards the time rises in proportion with the increasing number of event in the input stream.

6.1 Time consumption on Core nodes

The different times within CEP and DG namely *Start, Wait, Rules, Publish, Processing* and *Total* are measured and plotted on the graphs shown in Figure 6.2. The initialization phase of the nodes (*Start*) has been included here to show that it has no impact on the overall system performance (note: the number of rules and topics did not change).

An important metric evaluated is the time consumed during rule evaluation and the wait a node does before fetching the next event from the queue. These times, *Wait* and *Rules*, shown in Figure 6.3 and discussed in the following section, depict the performance of Drools Expert.

Another Important metric is the time each node takes to publish events to the topic. There are three nodes doing this task on their corresponding topics; the input console, CEP and DG. The publish time of these nodes measures the efficiency of RabbitMQ.

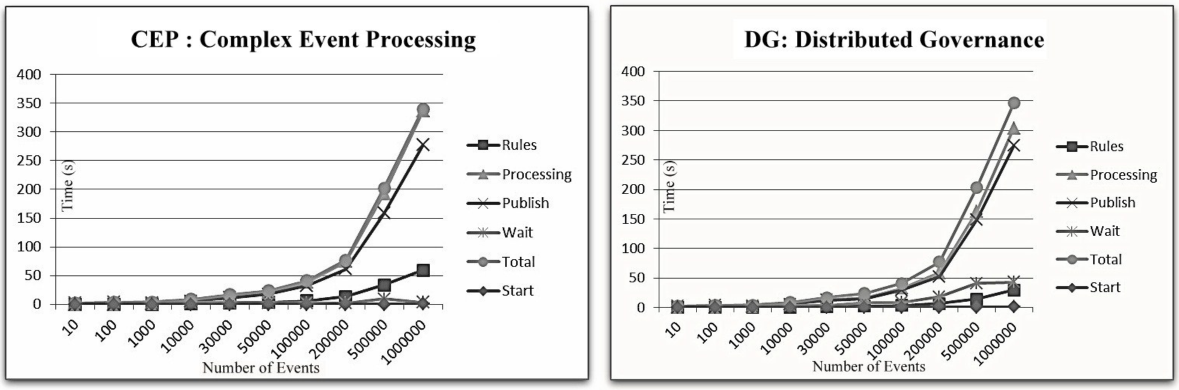


Figure 6.2: Different time consumptions in CEP and DG

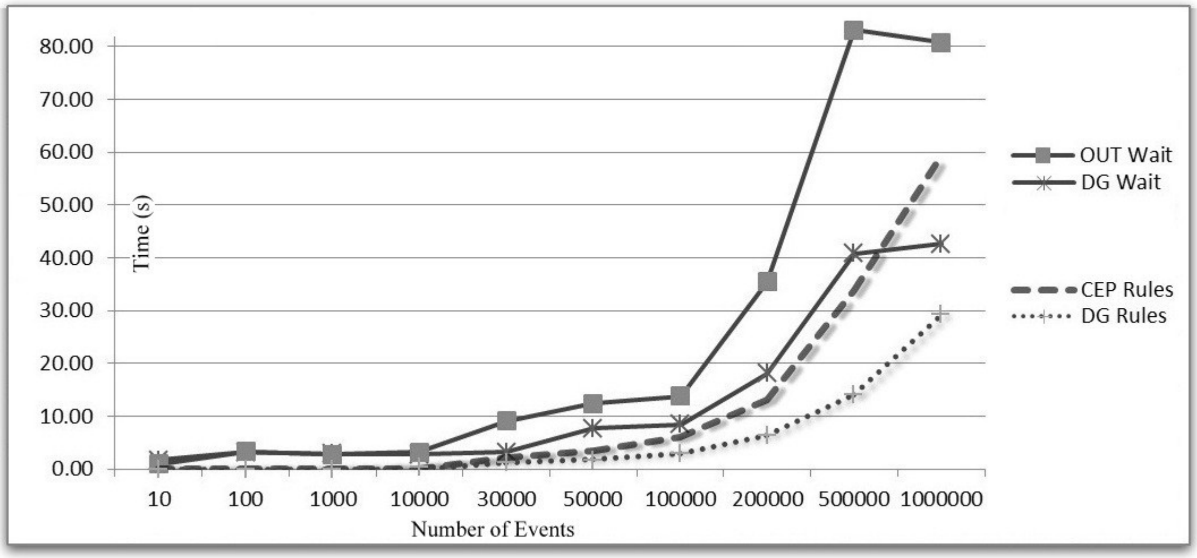


Figure 6.3: 'DG and OUT node Wait' vs 'CEP and DG Rules'

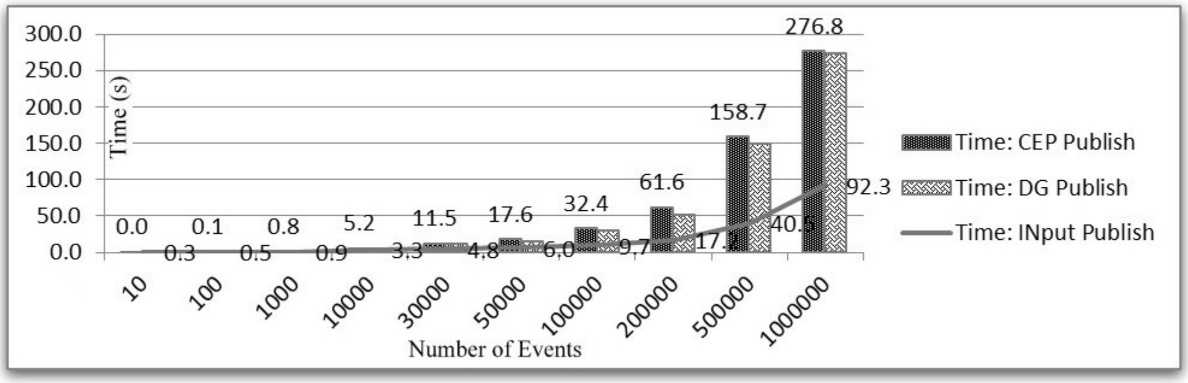


Figure 6.4: Publish time for Input, CEP and DG nodes

6.2 Discussion

Drools Expert rules are used on the nodes CEP and DG. DG's *Wait* is directly proportional to the time the CEP node takes for rule evaluation. The output console also waits with DG for CEP rules evaluation and then waits for the DG rules evaluation. Hence it has the longest wait time. Figure 6.3 compares wait time with the rules evaluation time with increasing number of events.

CEP rules are complex and identify patterns which takes more time compared to DG rules which are used to select actions associated to patterns.

There are three locations where events are publish: CEP, DG and the input console. Figure 6.4 shows the time it takes to publish. The time taken by the input console to publish all the events is very small (virtually negligible) for upto 30,000 events finishing even before CEP starts processing events.

Above 30,000 events, as the number of events increses it effects the CEP processing time and generates a cascading effect for the overall system performance. Thus, separating the input console from CEP (and subsequently DG) is important for any event stream above 30,000 events.

7 Summary and Future Work

This paper describes the first phase of our work on building a rule-based event processing system, which combines a messaging system with a rules system. We start by describing the underlying technologies, tools and products being used. Messaging using *AMQP* is implemented by RabbitMQ and our rule system uses Drools-Expert.

The architecture we have created consists of several interconnected components with communication links, realizing a distributed system. In our architecture we introduce 2 rule governing nodes; Complex Event Processing (CEP) and Distributed Governance (DG). We have streams of events entering the system which are being processed by CEP to generate complex events, essentially identifying patterns within the events. These complex events are then fed into DG for analysis and decisive action. The communication links between these components is provided by the messaging system. There are topic exchanges (channels and queues) between components which provide forwarding with selective filtering capability.

In this paper, we focused on the evaluation of performance of the products RabbitMQ and Drools by running several tests with events ranging from 10 to 1 million. In effect we are measuring the performance of rules and publishing of complex events. The *wait* state, introduced due to dependency of a node on processing time of previous node, is also considered.

Part of the future work planned is to deploy multiple CEP and DG nodes/engines that can work simultaneously to distribute the load at required times. We also have planned to increase the complexity of the governing rules in CEP and DG to test the highly complex patterns matching. A higher performance is the main objective of our work, currently we have all our nodes tested under constrained environment, working on a single machine (Intel i5, dual core) with Windows 7. Running our nodes across distributed servers in a cloud-based deployment should see the approach scale to a level appropriate for a high throughput, telecommunication grade management process.

References

[1] AMQP Architecture <http://www.amqp.org/architecture> (Last visited: 22/Feb/2013)

[2] RabbitMQ Tutorial <http://www.rabbitmq.com/tutorials/tutorial-one-java.html> (Last visited: 22/Feb/2013)

[3] RabbitMQ AMQP Concepts<http://www.rabbitmq.com/tutorials/amqp-concepts.html> (Last visited: 22/Feb/2013)

[4] JBoss.org <http://docs.jboss.org/drools/release/5.5.0.Final/drools-expert-docs/pdf/drools-expert-docs.pdf> (Last visited: 1/Feb/2013)

[5] Singh, J.; Vargas, L.; Bacon, J.; Moody, K.; *Policy-Based Information Sharing in Publish/Subscribe Middleware*. POLICY 2008. IEEE Workshop on Policies for Distributed Systems and Networks, Computer Lab., Univ. of Cambridge, Cambridge, pp.137-144, 2-4 June 2008

[6] Paschke, Adrian, and Alexander Kozlenkov. *A rule-based middleware for business process execution*. Multikonferenz Wirtschaftsinformatik. MKWI, 2008.

[7] Erradi, Abdelkarim, Piyush Maheshwari, and Vladimir Tosic. *Policy-driven middleware for self-adaptation of web services compositions*. Proceedings of the ACM/IFIP/USENIX 2006 International Conference on Middleware. Springer-Verlag New York, Inc., 2006.

[8] Geoffrey M. Winn, Neil G.S. Young. *Message oriented middleware with integrated rules engine* International Business Machines Corporation: US Patents US20130007184 (2012). <http://www.google.com/patents/US20130007184>

Indoor RTLS Using Smart Devices and Available Sensors

McGovern, John., McGrath, Kevin., Griffin, Leigh.

Telecommunications Software & Systems Group,
Waterford Institute of Technology, Waterford, Ireland
{jmcgovern, kmcgrath, lgriffin}@tssg.org

Abstract

Real Time Location System (RTLS) applications have the potential to open opportunities for developers to provide innovative new applications in areas as diverse as advertising, retailing and tourism sectors. Thus far however, these systems have been only able to specify Global Positioning System (GPS) as an interactive location provider or dependent on user interaction to physically 'check-in' their current location. The aim of this paper is to examine the implementation and testing of an RTLS developed specifically to use smart devices integrated with Bluetooth and Wi-Fi as the proximity sensors. This paper outlines the use of available sensors as proximity enablers, the algorithms needed for these sensors and how to test them within a building to identify Points Of Interest (POI). The expected granularity of each sensor and the potential pitfalls are also examined.

Keywords: Indoor Location, WiFi Triangulation, Bluetooth Triggering

1 Introduction & Related Work

Both ubiquitous [Kindberg, 2002] and pervasive computing [Satyanarayanan, 2001] which, were first introduced into the computing world in the 1990's, have generated a large amount of interest in the research world. However, to date they have seen a muted response from the wider commercial software development community and hence their introduction into mainstream computing. Cost has been a major contributing factor for this, requiring propriety hardware, specialised installation and needing propriety user devices. These devices also had the added disadvantage of requiring all potential end users to use the same operating handset device, which is only practical in a closed environment such as a factory setting. Reliable accuracy [Curran, 2011] of location in cheaper systems has to date been poor or sufficiently inaccurate to ensure a low market appeal.

With the proliferation of smartphones on the market today and the unprecedented uptake of these devices[Gartner, 2011], using smartphones to provide indoor Real Time Location Systems (RTLS) targeting smartphones as the context reader appeared an obvious next stage in pervasive awareness. Adopting a similar approach as that outlined by [Pfeifer, 2005], the team aimed to implement an indoor redundant positioning methodology running on smart devices as the receivers and to test the granularity of the location accuracy of this.

This software would allow application developers to use any available proximity data sensors to pinpoint a location without needing access to control servers. For instance in an indoor shopping mall the software would allow RTLS applications to infer location from specific shop Wi-Fis without needing access to the actual Wi-Fi Access Point (AP) or to maintain a central server of location data.

The following paper describes the sensor algorithms that were implemented, the testing techniques that were carried out and also the pit falls that the team encountered during this project. The paper is split into seven sections. Section 1 serves as the introduction. Section 2 describes the hardware components used during the development and testing stages. Section 3 describes the process needed to use the location software within a building. Section 4 describes the implementation and testing process for using Bluetooth as an input sensor. Section 5 describes the implementation and testing process for using Wi-Fi as an input sensor. Section 6 provides a discussion on the teams experience in trying to achieve low cost RTLS. Section 7 provides a conclusion and future work.

2 Infrastructure & Setup

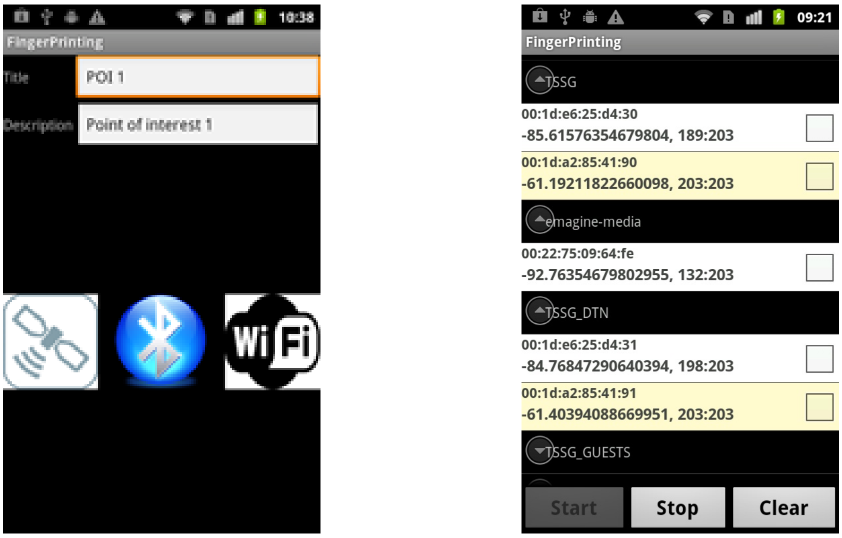
Development would require software for creating the test environment which is described in Section 3. Using experience gathered through [Pfeifer, 2009], the target building would be mapped or fingerprinted for each of the respective location sensors, this would in an RTLS suite of software become a tool of the administrative staff responsible for integrating a new location into the application offerings. A second effort would be required to develop a set of algorithms and application framework that would be the client or end user application area. This would incorporate the output from the fingerprinting and register a triggering of POI as the reference data matches the current sensor readings. The hardware used during development and testing was as follows:

- Mobile devices[McGoverna, 2013][McGovernb, 2013] running Android were chosen as the target platform as it was perceived these would offer the greater flexibility in available APIs.
- Bluetooth and Wi-Fi were chosen as the most widely available sensors for location reference points.
- PDA device [McGoverne, 2013] were used as the Bluetooth emitters. These can be plugged into a mains supply while emitting a Bluetooth signal this gave a great testing coverage without needing to recharge devices.
- Routers [McGovernf, 2013] were used during the library set up.

3 Finger printing

Buildings looking to utilise the software would require the proximity sensor data enabled, allowing them to be used as map determinants. To enable this an application was developed to run on the smart device allowing test locations to be mapped. At predetermined hotspots or POI an administrative user would be able to scan for available sensors and have the ability to select a set of these which would then act as a reference similar to that used by a topological map reference.

The administrative user responsible for mapping the building is presented with the option to select which sensor to use as part of the POI see Figure 1(a). They will then be allowed to scan for available devices under that respective sensor in range. A typical return from a WiFi scan is presented in Figure 1(b), the user can then select one or many of these that will then be used in the reference.



FingerPrinting

Title POI 1

Description Point of interest 1

FingerPrinting

TSSG

00:1d:e6:25:d4:30
-85.61576354679804, 189:203 □

00:1d:a2:85:41:90
-61.19211822660098, 203:203 □

emagine-media

00:22:75:09:64:fe
-92.76354679802955, 132:203 □

TSSG_DTN

00:1d:e6:25:d4:31
-84.76847290640394, 198:203 □

00:1d:a2:85:41:91
-61.40394088669951, 203:203 □

TSSG_GUESTS

(a) Available Sensors

(b) Available Wi-Fi Access Points (APs)

Figure 1: Available Connections

The POI mapping reference data is then saved to an Extensible Markup Language (XML) file on the device. This XML file was then used in the testing stages, however this would normally be bundled with the RTLS application software that would recognise the building POIs.

4 Bluetooth

Bluetooth was implemented as the first indoor location sensor, the algorithm developed would follow these basic steps:

- Poll to find available Bluetooth devices.
- Find their Received Signal Strength Indicator (RSSI) for each device found.
- Compare the RSSI to the fingerprinted model, if there was a match it is a successful POI.

The Bluetooth test environment was the development teams 9m x 6m office space depicted in Figure 2. Mio P550 PDA devices [McGoverne, 2013] were providing the Bluetooth signal emitters, the MAC address of each Mio device was determined and this was used to uniquely identify the sensor. Six devices were then placed in the room at head height. The Dell Streak 5 was used as the reading device. Both the Mio devices and the Dell were using the Bluetooth 2.0 protocols.

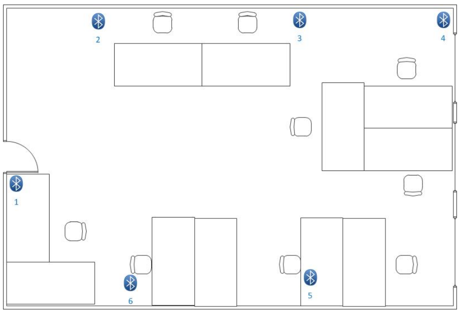


Figure 2: Bluetooth test environment.

The application measured the RSSI to indicate the power of the received signal at a given time. Testing was initially carried out using Bluetooth sensor 1, measuring the RSSI at successive one metre intervals, the results of four test cycles are presented in Figure 3. The initial testing indicated that using Bluetooth RSSI directly for <10 metre location accuracy was not a practical solution other than determining visibility.

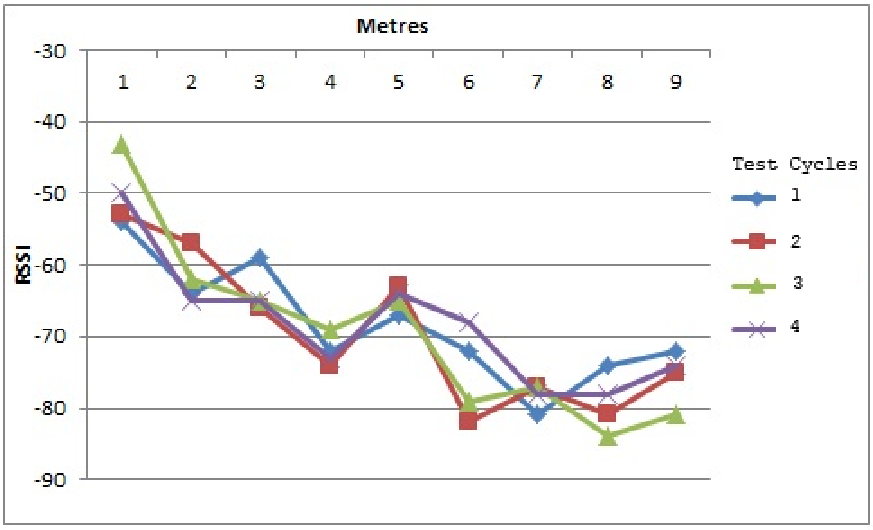


Figure 3: RSSI at one metre intervals.

Concluding that this wasn't a feasible RTLS solution the team adopted the approach of trying to narrow the focus of the signal similar to the home implementations of Wi-Fi signals [Flickenger, 2001]. The premise was to offer directional choice on the signal of each sensor, thus allowing positional choices in the room mapping. However it was found that this had no effect on the dispersion of the signal within the room.

Following on from this the team placed the sensor in a Faraday type cage (wrapping it in tinfoil to shield the signal from escaping) to examine if they could weaken the signal sufficiently to offer a narrower granularity on the detection. It was found that this could reduce the signal range to a 2-3 metre radius.

Following this success the team mapped 6 POI in the room and the tester walked randomly within the room in search of the POI. With some tweaking to the strength of the Faraday cage effect it was possible to accurately trigger each of the POI in multiple tests.

The next step to testing the Bluetooth algorithm was to ensure that it could scale to multiple devices, the results of which are outlined in Table 1. The scenario tested was how quickly could multiple devices simultaneously find the same sensor i.e. would a group of people arriving at the same POI at the same time have an impact. The detection algorithm was altered to run in a loop and testing was carried out on a range of devices from 1 to 15.

The 10.43 seconds average in the results from one device is due to the way that the current Android Java Bluetooth API implements its scan procedure i.e. it scans to find all devices and returns a list of these. This list can then be examined to see if the POI is in the list. It would have been faster to scan for a single sensor and detect if it was in range or not, however, the RSSI value is not returned if scanning is implemented in this manner.

The team experimented with other methods such as returning each found device as it finds them and examining if it is required. If this was one that was required the scan was stopped and returned the found POI, then recommenced a Bluetooth poll again. This had no significant impact on the results, it was hypothesised that the emitting sensor worked on its own timing loop equivalent to the 10.43 seconds found above however this was outside of our influence and this hypothesis was not confirmed at this time.

Number of Devices	1	3	5	7	10	15
Detection Attempts	250	750	1250	1750	2500	3750
Percentage of Fails	0	22.27	18.72	32.8	32.88	65
Avg. find Time (secs)[1]	10.43	11.74	12.81	16.07	16.18	37.18
Max find Time (secs) [2]	16.65	98.42	91.81	169.51	246.25	449.76

Table 1: Table of performance figures on scaling the BT polling.

5 Wi-Fi

Wi-Fi as a location mechanism has been introduced commercially by companies such as Ekahau [McGovernc, 2013] which utilises existing 802.11 Wi-Fi networks and Ubisense [McGovernd, 2013] which uses the Ultra Wide Band (UWB). Systems such as Ubisense require specialised hardware and tags to find the location data which are simply too costly for use in a non-specialised environment such as a large manufacturing workplace. Ekahau can use existing Wi-Fi technology but requires the smart device to connect to a positioning server and provide the collected RSSI data to the server, the server then determines the location and sends it back to the user. It was felt that to achieve redundancy as per [Pfeifer, 2005] that the Wi-Fi detection algorithms should run on the device in cooperation with the Bluetooth algorithms. An algorithm using the Euclidean distance formula was implemented.

$$d = |X - Y| = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2}$$

The complete office building was selected as the initial test environment. Wi-Fi analyser tests were ran in the building to give quick indicators on the choice of networks and their proliferation throughout the building. Up to 15 unique networks were visible within the building and all appeared to broadcast their MAC, SSID and RSSI values, which is the only information required. This formula is used to measure distance in a plane.

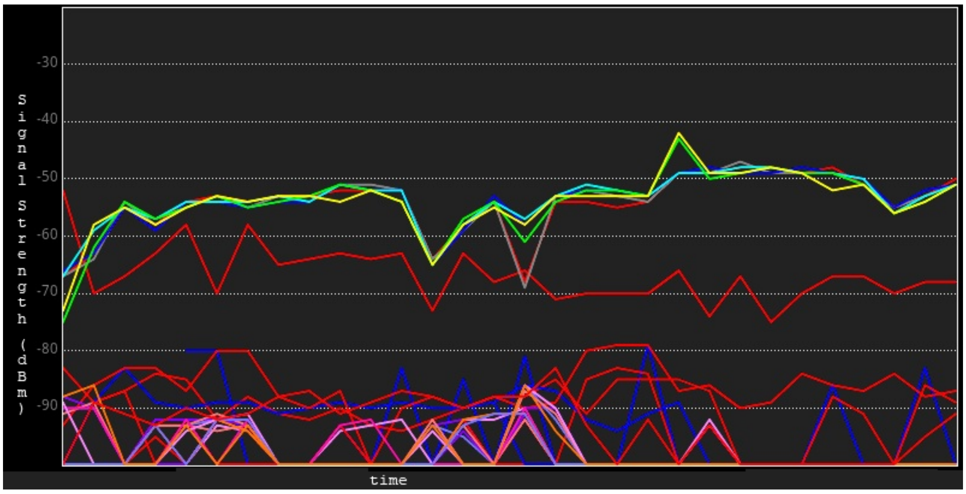


Figure 4: Sample Wi-Fi Analyser output from the office.

[1] Average find time: The length of time from poll initiation until Bluetooth polling completed.

[2] Max find Time: The greatest length of time a single device from the group took to finish a successful poll.

Fingerprinting of POIs was carried out and testing of the triggering process began. This proved to be a much more difficult task than was first anticipated. Firstly, while there was an excess of APs throughout the building they were created to provide data access and not configured for easy triangulation points. This gave the building a lot of dead spots where there wasn't a sufficient set of APs with a consistent signal to allow triangulation. A location within the building was found that would allow sufficient Wi-Fi reliability for initial testing to be carried out, however again results were inconsistent. The inconsistancy was mainly due to a normally good signal dropping momentarily, but long enough to give an inaccurate real time measurement. This inconsistency required changes to the algorithm namely to cover the following areas:

- A drop in signal was reported as a zero from the API. As can be noted from Figure 6, RSSI values are returned as negative numbers and a zero has a bias positive on the calculations.
- It was found that APs that provided a RSSI value of -75dBm or lower frequently failed to return any value as they could only be found intermittently so could not be included in any analysis, the algorithm needed to accommodate this by normalising the received RSSI.

Testing was moved to another location where the APs could be controlled in their location setting to allow for more comprehensive tests. The campus library was chosen, it provided a 12m x 14m open space. While a Wi-Fi network was in place in the library some initial tests suggested that only one AP would constantly produce an RSSI value of greater than -80dBm. The team added five additional APs and installed them strategically as shown in Figure 5 these were named simply as AP1 to AP5 for identification purposes.

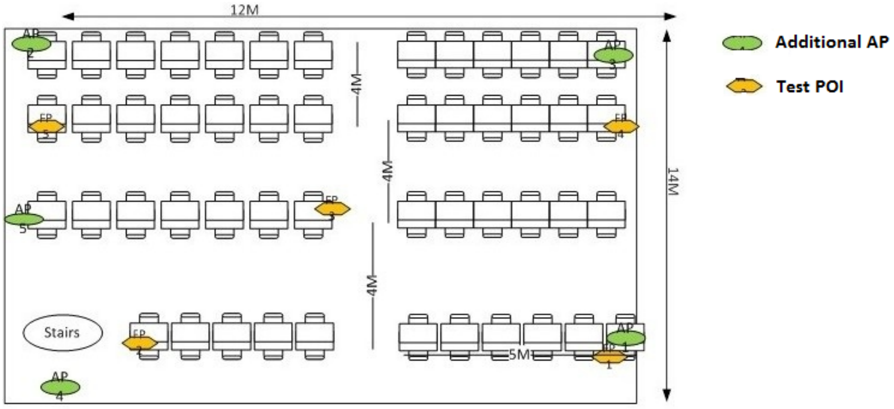


Figure 5: Library AP layout & POI testing positions

The room was divided into a logical 3 x 4 metre grid. The centre point of each grid was assigned as a unique POI and the location required for fingerprinting. The fingerprinting application was modified to take 250 RSSI readings per request at each of the POIs. These readings were then analysed statistically to determine if a detectible difference was discernible at each of the assigned POI. Table 2 & 3 show the average, minimum, maximum, median and standard deviation from the 250 RSSI values gathered at both POI 1 and POI 7 respectively.

This exercise was then repeated 5 times to ensure that the results were not anomalies and could be used as part of the triggering algorithm. The RSSI values received proved to be consistent across the range of testing.

Five of the POIs were then selected as the test POIs for the triggering test application, these are marked in Figure 5. The tester then walked randomly around the test site. It was proved that the Euclidean distance formula could be used to provide <5m accuracy with this configuration.

	AP1	AP2	AP3	AP4	AP5
Average	-49	-59	-65	-66	-69
Minimum	-59	-69	-74	-73	-71
Maximum	-46	-54	-61	-58	-66
Median	-50	-60	-66	-65	-68
Standard Deviation	2.58	2.35	2.04	2.41	1.10

Table 2: Fingerprinting POI 1

	AP1	AP2	AP3	AP4	AP5
Average	-69	-72	-45	-61	-66
Minimum	-72	-74	-50	-69	-71
Maximum	-65	-66	-40	-57	-64
Median	-68	-72	-47	-62	-65
Standard Deviation	0.969	0.969	3.342	1.662	1.914

Table 3: Fingerprinting POI 7

6 Results Discussion

To use **Bluetooth** as this paper describes to provide an RTLS system, implementation teams can expect two potential outcomes. At best, prove reliably that a person has entered a room. At worst, using Bluetooth with an average find time of 10.43 seconds and an estimated strolling pace of 1.3 m/s. An unlucky user could potentially mistime their arrival at a Bluetooth sensor and simply have walked past it before the scan results are returned.

When the expected number of people polling a specific device exceeds five, Bluetooth response slowed significantly Table 1 and its use as a sensor should be limited to situations where movement is restricted once you have arrived, i.e. you can walk past the sensor before you found it. Without more control on Bluetooth at both sensor side, in terms of its emitting loop and emitting range, and at device level, with the ability to poll a specific device list and return the RSSI. Bluetooth should not be considered as an option for reliable RTLS.

Using **Wi-Fi** it is the opinion of the team that unobtrusive Wi-Fi triangulation is possible in the shopping mall scenario as mentioned earlier with an expected 10 meter accuracy (unaided by additional APs) which would be an acceptable granularity given the space of a mall. However, factors such as building topology and AP positions had a major influence on the ability to utilise this technology easily. During testing the team found that Wi-Fi RSSI could be influenced by factors such as device location on or around the persons body relative to the APs and also the number of people in the room at a given time. Using Wi-Fi for fine granularity results was not a reliable location sensor, in fact it compounds this by demonstrating that the relatively small library space 168m2 used 5 APs to provide sub 5 meter location granularity.

7 Future Work & Conclusion

Attaining true redundancy [Pfeifer, 2005] as described did not happen, due to the time and effort taken during the Wi-Fi testing and configuration. More work will be carried out to introduce RFID algorithms to the sensor suite. Work in the area of predictive analytics will also be carried out that will aid in the detection of the next possible POI's that would under normal circumstances be triggered next.

Providing a reliable indoor RTLS that is capable of working within an existing building infrastructure will create a new breed of application development. This is an important area of research for innovative application development and economic development of the future.

This work is being carried out under the COIN project[1] which is investigating a broader pervasive environment.

References

[Kindberg, 2002] T. Kindberg, & A. Fox (2002). System Software for Ubiquitous Computing. *IEEE Pervasive Computing*, 1(1), 70-81.

[Satyanarayanan, 2001] M. Satyanarayanan. (2001). Pervasive computing: vision and challenges. *IEEE PCM*, 8(4), 10-17.

[Curran, 2011] Kevin Curran, EoghanFurey, Tom Lunney, Jose Santos, Derek Woods, Aiden McCaughey (2011). An evaluation of indoor location determination technologies. *Journal of Location Based Services*, Vol. 5, No. 2. (2011), pp. 61-78.

[Gartner, 2011] Gartner. (2011). Gartner's Hype Cycle Special Report for 2011.

[Pfeifer, 2005] T. Pfeifer. (2005). Redundant positioning architecture.. *Comput.Commun.* 28, 13 (August 2005), 1575-1585.

[McGoverna, 2013] J. McGoverna (2013). Dell Streak 5 Android 2.2. http://en.wikipedia.org/wiki/Dell_Streak, Last Accessed: 1922013.

[McGovernb, 2013] J. McGovernb (2013). Samsung Galaxy SII Android 2.3. http://en.wikipedia.org/wiki/Samsung_Galaxy_SII, Last Accessed: 1922013.

[Pfeifer, 2009] Tom Pfeifer, Paul Savage, and Bronwen Robinson (2009). Managing the culloden battlefield invisible mobile guidance experience.. In Proceedings of the 6th international workshop on Managing ubiquitous communications and services (MUCS '09). ACM, New York, NY, USA, 51-58.

[Flickenger, 2001] R. Flickenger. (2001). Antenna on the Cheap (er, Chip). http://www.oreillynet.com/cs/weblog/view/wlg/448.

[McGovernc, 2013] J. McGovernc (2013). Wi-Fi Based Asset Management and People Tracking Solution For Hospitals and Other Enterprises http://www.ekahau.com/products/real-time-location-system/overview.html, Last Accessed: 1922013.

[McGovernd, 2013] J. McGovernd (2013). Ubisense real-time location systems (RTLS) provide visibility of and control over assets and processes in multifaceted environments where objects are constantly on the move. http://www.ubisense.net/en/rtls-solutions/, Last Accessed: 19-22013.

[McGoverne, 2013] J. McGoverne (2013). The Mio P550 was a member of Mio's discontinued "Digiwalker" product line, and runs on Windows Mobile 2005, with Windows Media Player 10 for portables and portable Office. http://en.wikipedia.org/wiki/Mio_P550, Last Accessed: 1922013.

[McGovernf, 2013] J. McGovernf (2013). The Linksys WRT54G is a Wi-Fi capable residential gateway from Linksys. http://en.wikipedia.org/wiki/Linksys_WRT54G_series, Last Accessed: 1922013.

[1] Please see www.pervasiveengine.com for more details of this project

Deployment and Network Performance challenges for WSN based Environmental Monitoring Applications

Jacqueline Stewart, Robert Stewart, John Allen

Dept. of Electronics, Computer and Software Engineering,
The Wireless Networking Group
Athlone Institute of Technology,
Athlone, Co. Westmeath, Ireland.
{jstewart; jallen}@research.ait.ie
rstewart@ait.ie

Abstract

The ubiquitous deployment of IEEE 802.11 based Wireless Mesh Networks (WMNs) has facilitated non-real-time applications effectively to date. However advancements in technology in the area of Wireless Sensor Networks (WSNs) have seen a dramatic change in applications service requirements specifically real-time service deliver over heterogeneous wireless networks. Such networks often operate in hostile environments with limited line-of-sight and high interference levels. To cater for guaranteed Quality of Service (QoS) efficient traffic engineering tools such as network dimensioning and pre-deployment planning is essential.

The focus of this paper highlights the issues encountered in the deployment of the AIT-ECOMESH external environmental test-bed incorporating Crossbow WSNs utilizing a Motorola WMN as the backhaul network located within a dense woodland environment. Its location provides harsh environmental operational conditions for both networks in the area of RF signal propagation and interference mimicking those encountered in the precision agriculture industry were many new WSN applications are emerging.

This paper presents a range of empirical network performance results utilizing IX Chariot performance testing software. The results were statistically analysed and pre-planning recommendations are presented including the mitigation of near field effects, suggested limitations for deployment in dense vegetation and traffic characterisation to facilitate the dimensioning of networks. The performance measurements obtained will form the basis of future work to explore a new service level admission control protocol for real-time WSN applications which can ultimately be introduced by Industry to the customer as a guaranteed service agreement.

Keywords: Wireless Sensor Networks, Real-time Applications, Precision Agriculture, Signal Propagation

1 Introduction

The ever increasing advancements in technology and miniaturisation coupled with lower manufacturing costs have enabled and fuelled the current progress in new and exciting applications utilising WSNs. These small embedded system devices known as motes incorporate highly efficient wireless communication capabilities integrated with sensors capable of collecting data from the physical world and formulate the basic building block of each WSN. Each mote while working independently requires the service of other neighbouring motes to facilitate the transportation of its data over long distances. Independently these short range motes can reach a theoretical distance of between ~10-30 metres with a data rate of maximum 250Kbps. The network globally communicates through wireless links terminating at a gateway which further enhances its operational ability by forwarding data collected by the motes to a base station. The incorporation of a WMN as a backhaul network with the installation of mobile Access Points (APs) advances it's practical and cost efficient adaption for new applications. The acceptance of the IEEE 802.15.4 protocol as a communication standard for Low Rate Wireless Local Area Networks (LR-WLANs) together with the addition of the ZigBee specification (Zigbee Alliance 2005) defined a full protocol stack suitable to wireless sensors [1]. WSNs are described as a form of autonomous self-organized, self-healing, ad-hoc networks incorporating hundreds or thousands of low-rate motes which generally operate by battery power [2].

The topology of the WSN may change quite often to suit the application requirements and this positioning flexibility particularly suites the utilization of such networks in precision agriculture. Ecological agriculture which is also known as organic or biological farming involves the growing of crops while respecting the land and normally will not involve the use of chemicals such as pesticides or the use of genetically modified seeds. The verification of benefits achievable by the usage of wireless sensors is on-going in research facilitated in locations such as established experimental test-beds and large farms worldwide [3]. By reacting to environmental / climatic factors such as those found in vineyards like temperature, humidity, solar radiation, PH levels, potassium, sugar levels in grapes, water content and soil nutrients in real-time will result in optimizing production while establishing water saving policies, reducing production and labour cost, protect natural resources while mitigating the impact on the earth's environment [4] [5]. As all locations are not the same, deployment and operations of such networks within a hostile environment is providing researchers with various propagation, connectivity and QoS challenges.

A range of methodologies are available for deployment of a WSN in a hostile environment however due to the nature of the crops and terrain deployments vary enormously [6] [7]. The propagation of radio waves in such hostile environments are subject to major path loss degradation [6] [8]. The impact of vegetation on the communication channel is influenced by factors such as frequency, vegetation type, leaf state, vegetation density, weather conditions, water content, humidity and antenna height. Other factors such as ground reflection and canopy diffraction must also be incorporated in pre-deployment planning. The confirmation of ground reflection and canopy diffraction is highlighted later in this paper from experimental results recorded during testing at the ECOMESH test-bed. The difficulties in quantifying exact globally acceptable parameters have led to the creation and adaptation of different operational models [6].

1.1 Related Work

During the literature review process it was evident that many researchers have evaluated different vegetation attenuation models using Empirical and Analytical methods, while the majority prefer the use of Empirical methods due to its simplicity. Empirical models include Modified Exponential Decay (MED), Maximum Attenuation (MA), and Non-Zero Gradient (NZG) models [6]. Studies have shown that different parameter value results achieved from the usage of the MED model, has led to the development of variation models such as ITU-R, Weissberger, COST 235 and the FITU-R model. The Exponential Decay Model was the basic and first utilised model due to its simplicity.
Equation (1) below illustrates its format:

$$L(dB) = A\,x\,f^{B}d^{C} \tag{1}$$

Where A, B, and C are fitted parameters validated from experiments with regression techniques, such as frequency, foliage type, and propagation mechanisms [7]. Adaptations and developments of this basic model delivered numerous variations.

Analytical methods while more complex incorporate greater detailed input of the vegetation but are more computationally complex. Analytical methods include the Radiative Energy Transfer (RET) model which assumes the environment to be made up of a statistically even medium of scatters and absorbers. Detailed analysis of the various models is explored in [6] [7] [9] [8] however they focus on the attenuation of the signal and do not examine other network performance metrics.

In [8] a study of particular models operating within specified frequencies was explored. This research highlighted the lack of suitable collated experimental data available to research. It determined that at frequencies of 1GHz specific attenuation through trees in full leaf appears to be 20% greater (dBm) than when the trees are not in leaf. It was determined that a ground reflected wave component can cause signal propagation loss at a specified distance. Equation (2) below illustrated the loss experienced at the receiver due to the ground wave reflection [8]:

$$L_{ground} = 20\log_{10} \frac{d_1+d_2}{d_0} - 20\log_{10} R_0 + G_{Tx} \varphi + G_{Rx} \varphi \qquad (2)$$

where d_0 represents the distance direct from Tx to Rx, d_1 represents the distance from Tx to ground at an angle of θ_g while d_2 represents the distance of propagation from ground to RX. The reflection coefficient, R_0 of the ground reflected signal may be calculated with a given grazing angle θ_g while $G_{Tx} \varphi + G_{Rx} \varphi$ represent losses due to angles of the reflected wave from the transmit antenna then arriving into the receive antenna.

During the empirical experiments it was determined that ground reflection was evident at a distance of 30 meters. This analysis is later discussed in section 4. Other components also exist such as canopy diffraction wave (L_{top}) which is included below. Equation (3) illustrates the total loss L_{total} experienced by a signal propagating through trees derived by the combination of loss terms [8].

$$L_{total} = -10\log_{10} 10 \frac{-L_{sidea}}{10} + 10 \frac{-L_{sideb}}{10} + \frac{-L_{top}}{10} + 10 \frac{-L_{ground}}{10} + 10 \frac{-L_{scat}}{10} \qquad (3)$$

Experimental results and analysis presented in this paper incorporate seasonal variations in foliage density observed within the ECOMESH network. The results presented will inform research as to development of a best practice methodology for pre-deployment of WSNs utilizing a WMN as a backhaul system. It is the intention to utilize these measurements to form the basis of an admission control protocol for real-time WSN applications. A number of research groups have endeavoured to implement different priority mechanisms, resource admission control systems and service differentiation [10] [11] [12]. The research documented in this paper builds on the research to date of the AIT Wireless Networking Group as published in [13] [14] [15] [16].

The layout of this paper is as follows: Section 2 explores the ECOMESH test-bed system description and the methods for the collection of experimental data. Section 3 details the results over a ten month period. Section 4 presents an analysis of the results and delivers recommendations to industry in terms of pre-planning and deployment issues while Section 5 outlines the conclusion and future work.

2 ECOMESH Test-bed System Description

2.1 System Description

The ECOMESH test-bed was established in a dense woodland environment. Its deployment covers an area of 3-4 acres which is subdivided into three sections as per figure 1 below. The individual sections consisted of various degrees of foliage density and also included an open grass area. It abounds with different types of trees, mainly horse chestnut and whitethorn. The test-bed consists of off-the-shelf equipment. To the best of the author's knowledge there are no external environmental monitoring heterogeneous wireless test-bed deployments monitoring network performance currently in Ireland.

This research incorporated the use of Crossbow Iris Wireless Sensor Motes supplied by MEMSIC. These sensor motes are designed for embedded sensor networks with data rates of 250Kbps. They operate under the IEEE 802.15.4 standard at a frequency of 2.4GHz within the ISM band. This frequency incorporates 16 individual channels with a bandwidth of 3MHz each and each channel separated by 5MHz based on their centre frequencies. They are configured to transmit at a total power of 3.2dBm and typically have a receiver sensitivity of -101dBm. Each individual mote is powered by 2 AA size batteries. Attached to each mote is a data acquisition board or sensor board MDA100CB. This sensor board contains a precision thermistor, a light sensor/photocell and general prototyping area. The use of a MIB600 Ethernet Interface Board together provides Ethernet connectivity to the WSN for remote access. The use of a Symbol CB3000 bridge hardwired to the MIB600 effectively allows connectivity with the backhaul Enterprise Class Motorola WMN operating under the IEEE 802.11 standard. Remote access to the ECOMESH test-bed is established via remote access software "LogMeIn" which facilitates 24hr observation of experiments running within the test-bed.

2.2 Methods

The empirical experiments were conducted over 10 month duration. This was to facilitate the changes in foliage density within the test-bed during summer months with full leaf coverage on the trees and winter months with minimal leaf coverage. Figure 1 below illustrates the topology of the ECOMESH deployment.

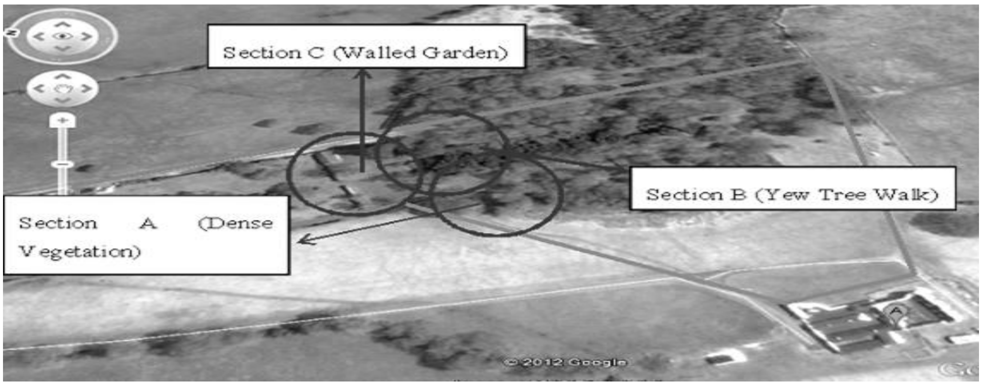

Section C (Walled Garden)
Section B (Yew Tree Walk)
Section A (Dense Vegetation)
©2012 Google

Figure 1: Topology of ECOMESH Test-bed (span 3-4 acres)

Figure 2 below illustrates the network architecture of the ECOMESH test-bed for experimental purposes. A deployment of 8 Crossbow motes were positioned at AP(1) and AP(2) separately.

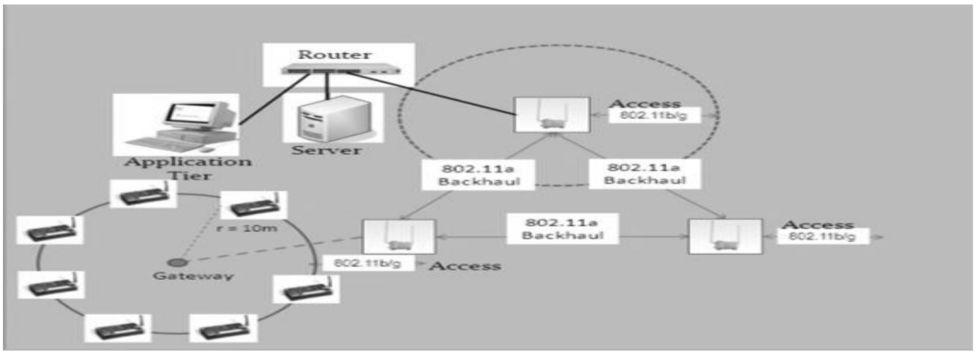


Figure 2: Network Architecture of Experimental Testing

Empirical experimental testing was carried out over one hop and two hops in relation to TCP and UDP traffic as per Figure 3(a) and Figure 3(b) below. Performance testing was further quantified by adding an eight mote WSN deployment to the backhaul network via AP1. Using IX Chariot software connected to the WMN utilizing the IEEE 802.11(g) standard at 2.4GHz various performance results were recorded. The configuration of the wireless connections between the individual APs operating under the IEEE 802.11(a) standard at a frequency of 5GHz completes the backhaul network. Various traffic parameters was explored in relation to distance, ground reflection and interference from the tree canopy by collecting the data at a distance of 10 metres from the AP and moving away until connectivity was lost. The height of the APs antennae was altered to examine the effect of canopy interference and that of the near field effect. These results were recorded and statistically analysed.

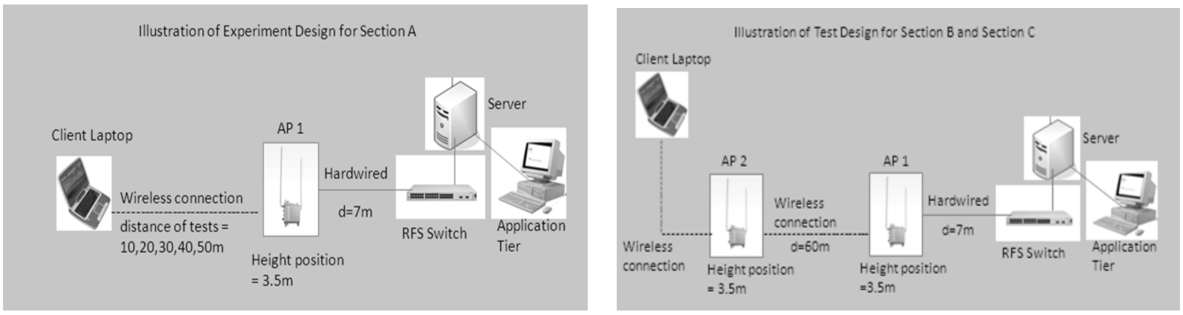


Figure 3(a) and 3(b): Illustration of Experimental Designs for Testing over one hop and two hops

Spectrum analysis and RSSI values were recorded utilising an R&S FSH 4/8 Spectrum Analyser. This determined the bandwidth of the signal and the signal power in dBm captured using Root Mean Square (RMS) power level. The recorded results incorporated a Signal to Noise Ratio (SNR).
Experimental performance results were compared with theoretical expectations.
The analytical/theoretical throughput is based on Equation (1) below:

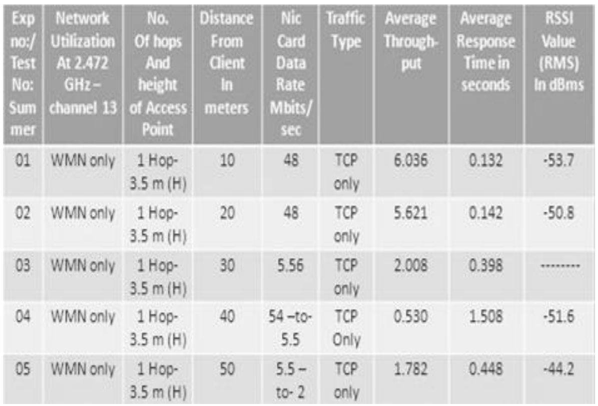
$$Throughput\ (Mbps) = \frac{1}{n} \tag{4}$$

Were *n* equates to the number of hops through the AP's.

Statistical analysis utilising Standard Deviation (STD) Matlab function was performed to evaluate environmental conditions, distance, weather conditions and operational loading within the network. This analysis determined the recommendations that forms the contributions stated in this paper.
Typically Internet type traffic is referred to as 'Bursty' traffic which can simply be described as traffic variability in terms of the amount of traffic generated in a given time frame. Studies have illustrated that even when aggregated applications are applied traffic smooth's out as more application traffic is multiplexed [17]. Experiment results provided evidence of smoothing discussed in section (4).

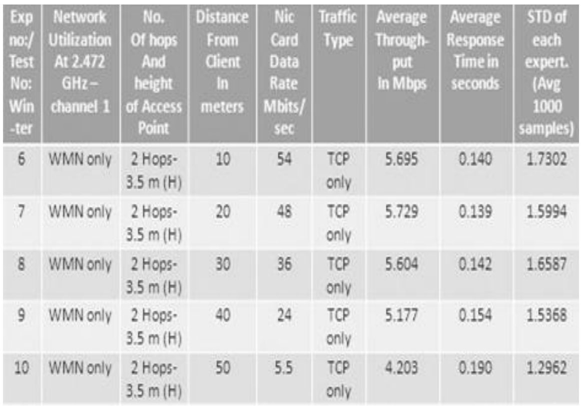
3 Results

The empirical results obtained are divided into summer test period and winter test period. The predominant vegetation was mature Horse chestnuts and Whitethorn. Table (1) and (2) below illustrates the performance measurements obtained during the summer and winter testing period:



Exp no:/ Test No: Summer	Network Utilization At 2.472 GHz – channel 13	No. Of hops And height of Access Point	Distance From Client In meters	Nic Card Data Rate Mbits/ sec	Traffic Type	Average Through-put	Average Response Time in seconds	RSSI Value (RMS) In dBms
01	WMN only	1 Hop- 3.5 m (H)	10	48	TCP only	6.036	0.132	-53.7
02	WMN only	1 Hop- 3.5 m (H)	20	48	TCP only	5.621	0.142	-50.8
03	WMN only	1 Hop- 3.5 m (H)	30	5.56	TCP only	2.008	0.398	--------
04	WMN only	1 Hop- 3.5 m (H)	40	54 –to- 5.5	TCP Only	0.530	1.508	-51.6
05	WMN only	1 Hop- 3.5 m (H)	50	5.5 – to- 2	TCP only	1.782	0.448	-44.2

Table (1): Summer Results over one hop



Exp no:/ Test No: Win -ter	Network Utilization At 2.472 GHz – channel 1	No. Of hops And height of Access Point	Distance From Client In meters	Nic Card Data Rate Mbits/ sec	Traffic Type	Average Through-put In Mbps	Average Response Time in seconds	STD of each expert. (Avg 1000 samples)
6	WMN only	2 Hops- 3.5 m (H)	10	54	TCP only	5.695	0.140	1.7302
7	WMN only	2 Hops- 3.5 m (H)	20	48	TCP only	5.729	0.139	1.5994
8	WMN only	2 Hops- 3.5 m (H)	30	36	TCP only	5.604	0.142	1.6587
9	WMN only	2 Hops- 3.5 m (H)	40	24	TCP only	5.177	0.154	1.5368
10	WMN only	2 Hops- 3.5 m (H)	50	5.5	TCP only	4.203	0.190	1.2962

Table (2): Winter Results over two hops.

Each experiment ran for two minutes in duration with client/laptop antenna positioned 0.5 meters from ground level and the AP antenna at 5 meters from ground level. Validation of the selection of two minutes duration per test was confirmed by the duplication of a second test ran over fifty minutes duration which returned a result of 5.502Mbps compared to 5.987Mbps, a difference of 0.485Mpbs. Figure 4(a) TCP traffic and Figure 4(b) UDP traffic illustrate the STD results obtained using throughput performance results over one hop commencing at 10 meters distance from the AP progressing to 70 meters. The test-bed section under test was open field.

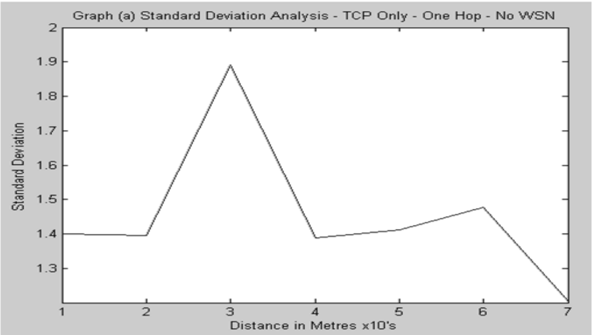


Figure 4(a): STD-TCP -one hop in winter

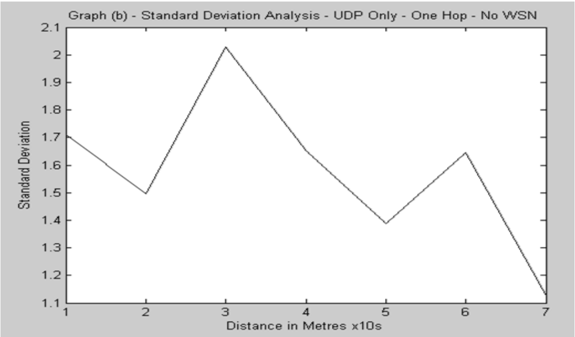


Figure 4(b): STD-UDP -one hop in winter

Figure 5(a) illustrates STD analysis comparing TCP throughput over one hop in open field during the winter season. It illustrates STD results between a deployment of an eight mote WSN and no WSN. Figure 5(b) illustrates STD using UDP traffic again comparing results between an eight mote WSN deployment and no WSN increasing at intervals of ten metres. The client/laptop antenna height position was at 0.5 m.

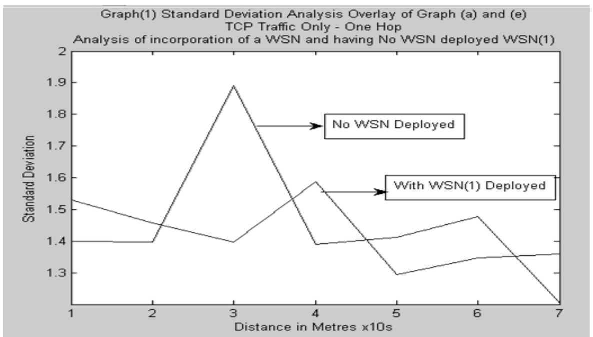


Figure 5(a): STD Analysis of TCP traffic

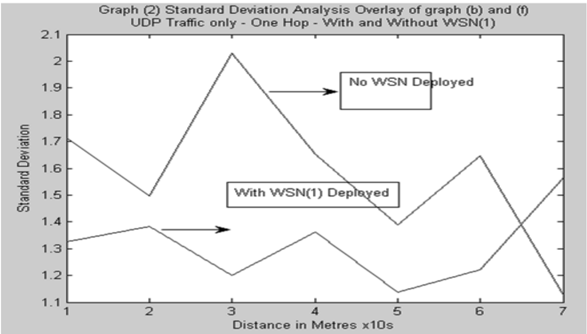


Figure 5(b): STD Analysis of UDP traffic

Figure 6 contains the results of Throughput analysis for TCP traffic with no WSN deployment which illustrates the coverage determined between the open field section and the dense vegetation section.

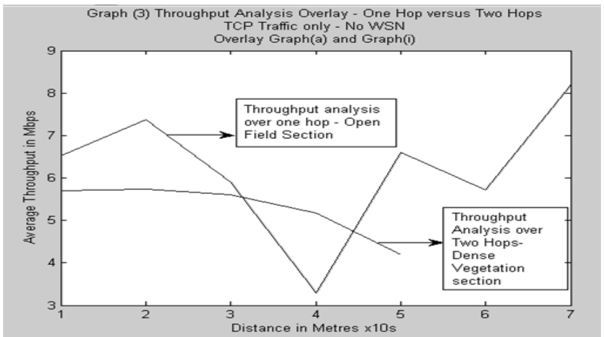


Figure 6: Comparison Throughput Analysis between open field and dense vegetation –No WSN deployed.

4 Discussion

Comparing the TCP throughput results taken in the section of dense vegetation, Table (1) and Table(2) above, it was determined that the throughput during winter increased by 65% compared with those of summer. This increase was achieved despite the link incorporating two hops where performance levels were expected to drop by 1/n. This illustrates the propagation challenges as per Equation (3) above experienced when vegetation is at its highest in full leaf in the summer.

Observation of STD results as per Figure 4(a) and Figure 4(b) illustrate a value of between 1.8 and 2.1 at a distance of 30 metres from AP(1). Both tests ran concurrent with AP antenna height (5 meters) and client antenna (0.5 meters) however the traffic type differed between TCP and UDP. The analysis of these results indicates that at 30 meters the propagation signal suffers from ground wave reflection as per Equation (2) above. Further testing is planned to validate results by running experiments with AP antenna height positioned at intervals of 1 meter from ground level to a maximum of 10 meters.

Analysis captured in Figure 5(a) and Figure 5(b) illustrates STD for TCP traffic and UDP traffic comparing results between no WSN deployment and the inclusion of an 8 mote WSN. The effect of the introduction of a WSN illustrates 'smoothing 'of the bursty characteristics of the traffic with a larger effect on the UDP traffic decreasing the STD to between 1.2 to 1.4.

Comparing the analysis of throughput, reliability and connectivity results recorded during the summer testing period and the winter testing period determines a limit on coverage attainable when propagation is in dense vegetation. Test results in the open field gave throughput averages for TCP traffic of 8.192 Mbps at a distance of 70 m as per Figure 7 above, while the throughput average recorded in dense vegetation was 4.203 Mbps at a distance of 50 m from the AP. Connectivity was lost at a maximum distance of 50m.

The importance of validation of theoretical analysis utilizing empirical experimental testing is evident by the results obtained from the external ECOMESH test-bed. As real-time applications require limitations on delay to facilitate the real-time actions of actuators increase, QoS is essential. Seasonal factors such as foliage density require inclusion in the pre-planning stage as results from the ECOMESH determined a 65% average fall in throughput due to increase in foliage density.

5 Conclusion

This paper has presented important challenges requiring consideration to aid in dimensioning of networks for real-time WSN applications in hostile environments similar to that of Agri-tech industry. The objectives of the various tests carried out over a 10 month period was to collect detailed measurements in relation to signal throughput, connectivity, distance and difficulties present in the propagation of such signals. The empirical results detailed in relation to the limitations of connectivity to a distance to 50 meters in dense vegetation, the presence of ground reflection at 30 meters in open field, the effects of smoothing with the addition of an 8 mote WSN and the evaluation of seasonal throughput to determine an increase of 65% during the Winter months have been detailed highlighting the importance of measurements in dimensioning of WSN's.

5.1 Future Work

Continued evaluation of the ECOMESH through experimental testing is planned to further validate the near field effect while incorporating other performance issues. The addition of an Industrial standard environmental monitoring camera to the test-bed is near completion with network performance analysis commencing shortly and re-evaluation during the summer months. The empirical performance measurements will form the basis of future work in relation to an Admission Control algorithm to quantify parameters for a guaranteed QoS delivery to the end user with further research developed using an OPNET^{TM} simulation model.

References:

[1] A. Koubaa, M. Alves and E. Tovar, “IEEE 802.15.4: a Federating Communication Protocol for Time-Sensitive Wireless Sensor Networks,” in *Sensor Networks & Configurations: Fundamentals, Platforms & Experiments,* Germany, Springer-Verlag, 2007, pp. 19-49.

[2] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam and E. Cayirci, “Wireless Sensor Networks: a survey,” *Computer Networks,* vol. 38, pp. 393-422, 2002.

[3] J. Lopez Riquelme, F. Soto, P. Suardiaz, P. Sanchez, A. Iborra and J. Vera, “Wireless Sensor Networks for precision horticulture in Southern Spain,” *Computers and Electronics in Agriculture,* no. 68, pp. 25-35, 2009.

[4] S. Diaz, J. Perez, A. Mateos, M. Marinescu and B. Guerra, “A novel methodology for the monitoring of the agricultural production process based on wireless sensor networks,” *Computer and Electronics in Agriculture,* no. 76, pp. 252-265, 2011.

[5] J. Huircan, C. Munoz, H. Young, L. Von Dossow, J. Bustos, G. Vivallo and M. Toneatti, “ZigBee-based wireless sensor network localization for cattle monitoring in grazing fields,” *Computers and Electronics in Agriculture,* no. 74, pp. 258-264, 2010.

[6] D. Ndzi, L. Kamarudin, E. Mohammad, A. Zakaria, R. Ahmad, M. Fareq, A. Shakaff and M. Jafaar, “Vegetation Attenuation Measurements and Modeling in Plantations for Wireless Sensor Network Planning,” *Progress In Electromagnetics Research B,* no. 36, pp. 283-301, 2012.

[7] Y. Meng and Y. Lee, “Investigations of Foliage Effect on Modern Wireless Communication Systems: A Review,” *Progress in Electromagnetics Research,* no. 105, pp. 313-332, 2010.

[8] I. T. U.-R. Sector-, “ITU-R P.833-7 Attenuation in Vegetation,” international Telecommunication Union, Geneva, Switzerland, 2012.

[9] G. Mao, B. D. Anderson and B. Fidan, “Path Loss exponent estimation for Wireless Sensor network localization,” *Computers Networks,* no. 51, pp. 2467-2483, 2007.

[10] B. C. Villaverde, S. Rea and D. Pesch, “D-SeDGAM: A Dynamic Service Differentiation Based GTS Allocation Mechanism for IEEE 802.15.4 WSN,” in *2010 7th International Conference on Information Technology: New Generations (ITNG),* Las Vegas, NV, USA, 2010.

[11] Y.-K. Huang, A.-C. Pang and H.-N. Hung, “An Adaptive GTS Allocation Scheme for IEEE 802.15.4,” *IEEE Transactions on Parallel and Distributed Systems,* vol. 19, no. 5, pp. 641-651, May 2008.

[12] A. Melikov and A. Rustamov, “Queuing Management in Wireless Sensor Networks for QoS Measurement,” *Wireless Sensor Networks,* vol. 4, pp. 211-218, 2012.

[13] J. Stewart, R. Stewart, M. Hassan and J. Allen, “Application of Decay Rate Analysis for GTS Provisioning in Wireless Sensor Networks,” in *The 8th IEEE International Symposium on Communications Systems, Networks and Digital Signal Processing.*, Poznan, Poland., July 2012.

[14] J. Stewart, R. Stewart, M. Hassan, A. Md.Shakaff and D. L. Ndzi, “Real Time Service Provisioning on the AIT Integrated Mesh Network for Precision Agri-tech Applications,” in *The IT&T 11th International Conference on Information Technology and Telecommunication* , Cork, Ireland., March 2012.

[15] J. Stewart, R. Stewart and M. Hassan, “The Application of Decay Rate Analysis for WSN Buffer Dimensioning,” in *Wireless Telecommunication Symposium,* London, UK., March 2012.

[16] Hassan: MohdNajmuddinMohd, “Improving Quality of Service for Real-Time Applications in Wireless Sensor Networks,” Athlone Institute of Technology, Athlone, Co. Westmeath, Ireland, 2012.

[17] P. Tinnakornsrisuphap, W. Feng and I. Philp, “On the Burstiness of the TCP Congestion-Control Mechanism in Distributed Computing System,” in *The 20th International Conference on Distributed Computing Systems (ICDCS 2000),* Taipei, Taiwan., 2000.

Poster Session – Short Papers

Opening the Door to Innovation in Cloud Computing

Trevor Clohessy [1] and Thomas Acton [1]

[1] ***Business Information Systems & Lero, Whitaker Institute,***
National University of Ireland Galway, Ireland.

{t.clohessy2, thomas.acton}@nuigalway.ie

Abstract

This paper describes research-in-progress that explores the applicability and implications of cloud computing in the creation of business value through open innovation. Both the cloud computing and open innovation paradigms represent recent phenomenon and as such many unanswered questions still persist. In responding to this research gap we propose a new value creation framework which is based on a review of the literature on cloud computing, innovation, open innovation and value. Taking the framework layer by layer, this research in progress shall evaluate the innovation potential across components capable of offering value to organisations. The main contribution of this paper lies in proposing a framework that seeks to identify best route(s) to value, thus providing a visual mapping to enable organisations determine which cloud computing components, implementations, solutions and innovation approach is most suitable for value attainment.

Keywords: Cloud computing, Innovation, Open Innovation, Value

1 Introduction

While some research has been carried out in order to determine how organisations can reap the benefits associated with cloud computing e.g. (Armbrust et al., 2010; Brynjolfsson, Hofmann, & Jordan, 2010; Buyya, Yeo, Venugopal, Broberg, & Brandic, 2009; Weinhardt et al., 2009), there is no empirical study which has examined how the principles of open innovation could complement a cloud computing approach for the creation of value. Nor has research looked at how individual components of the cloud computing model layers are more conducive than others for the attainment of value. Thus it is the objective of this study to explore the notion of cloud computing, its applicability, implications in a multiple partnering project ecosystem in order to identify key cloud centric enablers of value and ascertain the model of innovation utilised in the process.

2 The Conceptual Framework

For our theoretical base, we propose a layered 5-4-3-2-1 framework model. The Mell and Grance (2010) definition of cloud computing is specific in delineating the cloud as comprising five essential characteristics, four deployment models, and three service models. It is this definition and delineation that we employ in this paper, in particular, what we term the 5-4-3 cloud computing stack model layers comprising *the essential characteristics layer*, *the deployment model layer* and the *service model layer* (Clohessy & Acton, 2013). To capture the concepts of open and closed innovation, we consider an *innovation layer* composed of two innovation sub-layers, open and closed. However, we propose that there are more pathways to value creation through openness in innovation in contrast to a closed innovation approach. As a consequence, the open innovation layer is awarded a more prominent visual sizing in our framework model. The framework model (see figure 1) depicts a potential pathway to value. The model provides a useful lens with which to identify key cloud-centric enablers of business value and the method of innovation utilised in the process of its attainment. Taking the framework layer by layer, this research in progress shall delineate the innovation potential across components.

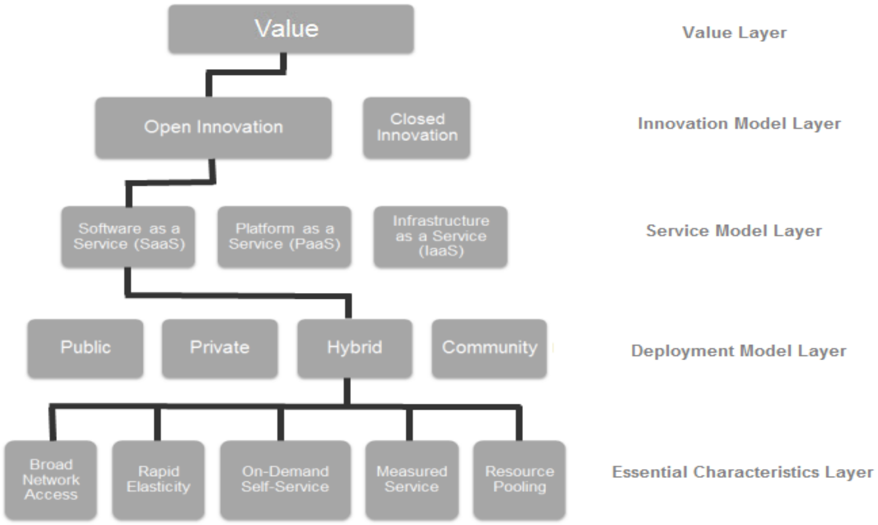


Figure 1: Framework Model (the 5-4-3-2-1 representing the number of elements at each layer)

3 Case

The initial case is a multinational information technology corporation who provide technology and software solutions to consumers and enterprises. The corporation also provide a cloud partner ecosystem which permits their partnering organisations to enhance their own cloud offerings in an attempt to attract new customers and gain entry into new markets. The cloud partner ecosystem consists of a set of tools, documentation, support and best practices provided by the case corporation. In this study we propose to analyse five partnership projects within the case corporation. The profiles of the five projects shall differ in terms of the service model and deployment model utilised.

4 Future Steps

This paper outlined research in progress aimed at exploring the applicability and validity of utilising cloud computing as a means of attaining value through open innovation in a multiple-project environment. Each project will be examined in the context of the cloud computing innovation framework. The study will provide insight into the innovation value of discrete components of the conceptual framework that facilitate the creation of an innovation pathway through the model, a ‘visual map’, that an organisation may traverse in order to facilitate the attainment of value.

Acknowledgements

This work was supported, in part, by *Science Foundation Ireland* grant 10/CE/I1855.

References

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A. & Zaharia, M. (2010). A View of Cloud Computing. *Communications of the Acm*, 53(4), 50-58.

Brynjolfsson, E., Hofmann, P., & Jordan, J. (2010). Cloud computing and electricity: beyond the utility model. *Communications of the Acm*, 53(5), 32-34.

Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation Computer Systems*, 25(6), 599-616.

Clohessy, T., & Acton, T. (2013). Value Creation and Capture with Cloud Computing. Paper presented at the Proceedings of the 18th Annual Conference of the UK Academy of Information Systems, Oxford, UK.

Mell, P., & Grance, T. (2010). The NIST Definition of Cloud Computing. Communications of the Acm, 53(6), 50-50.

Weinhardt, C., Anandasivam, A., Blau, B., Borissov, N., Meinl, T., Michalk, W., & Stosser, J. (2009). Cloud Computing - A Classification, Business Models, and Research Directions. *Business & Information Systems Engineering*, 1(5), 391 - 399.

A Hybrid Vehicular Re-routing Strategy with Dynamic Time Constraints for Road Traffic Congestion Avoidance

Shen Wang [1], Soufiene Djahel [2] and Jennifer McManis[1]
[1] Lero, RINCE, School of Electronic Engineering, Dublin City University, Ireland
shen.wang4@mail.dcu.ie, mcmanisj@eeng.dcu.ie
[2] Lero, School of Computer Science and Informatics, University College Dublin, Ireland
soufiene.djahel@ucd.ie

Abstract

Intelligent Transportation System (ITS) provides a promising framework to alleviate the congestion on the roads, but it still needs to be improved , such as in the area of vehicles re-routing strategies. The main focus of this paper is on designing novel vehicles re-routing strategy driven by dynamic time constraints to reduce the traffic congestion in urban areas. The next step of our work is to evaluate the performance of our strategy and compare it with the existing algorithms.

Keywords: ITS, Vehicles Re-routing, Shortest Path, Heuristic Algorithms.

1 Introduction

Smart routing of vehicles is one of the key services offered by ITS for achieving optimal load balance of the traffic on the roads. Most of the existing commercial routing products cannot react to sudden changes in route conditions during the journey. Therefore, real-time re-routing strategies, which can help the vehicles to react to any update in traffic conditions, including incidents, are highly required. In order to achieve this goal, upon advertisement of a real-time event, such as vehicle crash, stalled vehicle on the road, congested road segment etc, the re-routing algorithms should find an alternative optimal/near-optimal route before the vehicle reaches the next junction; otherwise it will probably be difficult to avoid the congestion.

In this paper, we propose a new vehicles re-routing strategy driven by dynamic time constraints. These time constraints refer to the remaining time for a vehicle to reach the next junction.

2 Related works

The vehicular routing problem is a variant of the classical shortest path problem, with the link's(road segment) weight should be either travel distance or travel time. Exact algorithms provide the best route according to the specified criteria and link weights. Nevertheless, there are no guarantees on their response time. Dijkstra Algorithm (DA) [1] is the most typical algorithm in this category. It searches the shortest route from one node to all other nodes in the road network. Moreover, it guarantees the termination property. Its time complexity is O(V^2), where V is the number of vertices (junctions). Due to the large size of real road networks, it cannot ensure short response time. For Heuristic Algorithms (HA), Hitoshi Kanoh [2] applied the virus genetic algorithm (GA) to dynamic route planning. Horst Wedde et al. [3] introduced a new system called Bee Jam Avoidance (BJA) inspired by the form of bee foraging communication. but none of them is able to react efficiently and within short time threshold to the potential change in road conditions during the vehicle journey. This is because they use a fixed computational time threshold which is less-efficient and unrealistic. Additionally, no performance comparison among them is conducted and most of them have been evaluated against DA only.

3 Our proposal

To overcome the drawbacks of the aforementioned solutions, we propose a hybrid vehicle route planning strategy based on proactive real-time event report mechanism. As shown in Figure 1, at the initialisation stage of our strategy, the road network map is loaded along with real time data and the corresponding prediction information (e.g. estimated average speed of vehicles in a certain road segment, estimated traffic congestion level after X minutes, etc). Simultaneously, the driver should input the desired destination, the vehicle features (e.g. type, height, size, emission of engine etc.) and route planning metrics (travel time, easiness of driving or fuel consumption level) to the system. Then, the driver gets the best route to destination using DA before the journey starts. While the vehicle is running on the road, if it receives real-time event report, the map with real-time data should be updated. Notice that for the whole hybrid re-routing system, the report mechanism works as an interruption in computer architecture; it can push the event information to the related vehicles with highest priority pro-actively, rather than letting the vehicles check the related status periodically. We assume that the vehicle will use the initial route plan until its destination is reached, if there is no real-time event received during the entire trip. If an event is received, according to the comparison between the computation time of DA (T_{DA}) and current due time (i.e. estimated travel time from current position to next intersection) T_d, the algorithm uses DA or HA to provide the new optimal route. In this case, the faster HA can be implemented. The idea behind this design is to make full use of the dynamic due time. If T_d is long enough, we do not need HA which can only provide near-optimal solution instead of the optimal (best) one. Finally, the algorithm ends up with a sign of destination road segment reached.

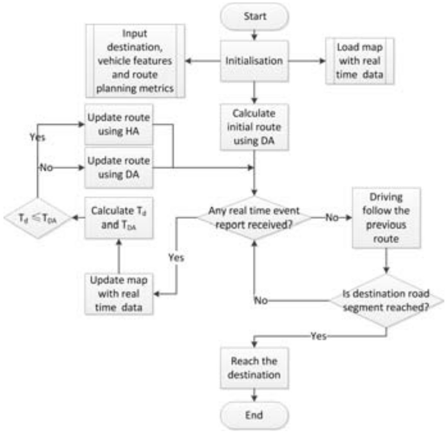

Start
Input destination, vehicle features and route planning metrics
Initialisation
Load map with real time data
Update route using HA
Calculate initial route using DA
Yes
No
Update route using DA
$T_d \leq T_{DA}$
Calculate T_d and T_{DA}
Any real time event report received?
No
Driving follow the previous route
Yes
Update map with real time data
No
Is destination road segment reached?
Yes
Reach the destination
End

Figure 1: Flowchart of our Hybrid Re-routing Strategy

We use the following metrics to evaluate performance of our strategy and existing solutions: travel time, easiness of driving, computation time and scalability. The travel time is crucial as the short travel time means high work productivity and more satisfaction of the city's citizens, everyone in this society can thus benefit from this improvement. The factor easiness of driving matters because if the drivers do not get comfortable driving experience from a system, they will choose another product. Moreover, if they drive in their easiest way the risk of their involvement in accidents will decrease. We can score the difficulty of the planned route in terms of number of turns, traffic lights, and the number of lanes in each road etc. As the re-routing solution has to be provided to the vehicles before they reach the next intersection, hence the computation time is important as well. Finally, we will take the scalability into account as the algorithm efficiency would be totally different when it is applied to various map layouts or different traffic loads. To implement the re-routing strategies, we will use the most widely used open-source microscopic transportation simulator named SUMO [4] and we use TAPASCologne [5] as real traffic flow.

References

[1] E. W. Dijkstra. A Note on Two Problems in Connection with Graphs. Num. Mathematik, 1:269-271, 1959.

[2] H. Kanoh, Dynamic route planning for car navigation systems using virus genetic algorithms, International Journal of Knowledge-based and Intelligent Engineering Systems, vol. 11, pp. 65-78, Jan. 2007.

[3] S. Senge and H. Wedde, Bee inspired online vehicle routing in large traffic systems, in ADAPTIVE 2010, The Second International Conference on Adaptive and Self-Adaptive Systems and Applications, 2010.

[4] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, SUMO - Simulation of Urban MObility - an Overview, in SIMUL 2011, The Third International Conference on Advances in System Simulation,(Barcelona, Spain), 2011.

[5] Uppoor, Sandesh, and Fiore Marco. "A large-scale vehicular mobility dataset of the Cologne urban area." 14mes Rencontres Francophones sur les Aspects Algorithmiques des Tlcommunications (AlgoTel) (2012): 1-4.

Adaptive Artificial Intelligence Utilizing Multi Modal Emotion Recognition

Alan Murphy [1], Sam Redfern [2]

[1] Department of Information Technology,
National University of Ireland Galway
a.murphy30@nuigalway.ie

[2] Department of Information Technology,
National University of Ireland Galway
sam.redfern@nuigalway.ie

Abstract

This research aims to apply a machine learning middleware to the development process of mainstream computer games. Every game has an intended purpose and intends to elicit a particular emotional response from its players. Monitoring the emotions of a player during game-play allows developers to assess whether their game is fulfilling its intended purpose. We propose middleware that can be used in the beta testing stages of a game's development, to extract user information from players, which can in turn be used to adapt game-play to elicit the desired emotion.

Keywords: Adaptive Artificial Intelligence, Emotion Recognition

1. Introduction

Scripted intelligence in modern games has a dramatic effect on a player experience during game-play. Park et al. (2012) outline how playing either with or against scripted NPCs and the level of intelligence of the NPCs can have a profound impact on the enjoyment levels of the player characters. Research until now has had little focus on the players' reactions to these enemies. Such a perspective could provide game developers with the user generated player data they require in order to script the AI for such enemy NPCs in order to elicit the expected emotional response from a game's players. This paper will initially outline the motivation for the work to be carried out, and will then proceed to outline the structure and setup of our prototype recognition system.

2. Emotion Recognition Setup

Our system will classify emotion based on a voice signal and also text obtained via a voice-to-text converter. What this provides is two systems attempting to solve the same problem and, therefore, the potential for increased accuracy. For our prototype recognition system a standard feed forward neural network with back-propagation training is being used. This network will process the audio data from players during game-play. The basic structure takes the form of a standard feed forward neural network and consists of three separate layers: an input layer; a hidden layer where our classification is performed, and a final output layer with six outputs to represent six fundamental emotions.

In terms of selecting audio parameters to use as input to our network for initial training, there is much research literature which endeavors to derive the optimum set of parameters for emotional recognition through acoustic analysis. Schuller et al. (2009) perhaps provides the most informed and complete parameter list which should encompass all acoustic aspects of the emotions conveyed. This list suggests the following seven acoustic parameters, and we use these as our base seven parameter:

Raw Signal:	Zero-crossing-rate
Signal energy:	logarithmic
Pitch:	Fundamental frequency F0 in Hz via Cestrum and Autocorrelation Exponentially smoothed F0 envelope
Voice Quality:	Probability of voicing
Spectral Energy:	Energy in bands 0-250 Hz, 0-650 Hz, 250-650 Hz, 1-4 kHz
Mel-spectrum:	Band 1-26
Cepstral Coefficients:	MFCC 0-12

Through experimentation we found that the sampling rate of the input audio stream has a direct influence on the pitch extracted from the signal. By using a higher sampling rate of 48 kHz this ensures that the pitch reading is indeed accurate, and for lower sampling rates of 16 kHz pitch readings were obscure. The reason for this is that in order to calculate the pitch parameter we must rely on higher harmonics of F0. In the 48 kHz signal there are more of these harmonics than in a 16 kHz signal, so this implies that pitch extraction will be more robust using the higher sampling rate. Currently our work is focused around building the appropriate input schema to train our neural network. This in turn informs our work in further deriving the seven parameters as outlined by Schuller et al.

3. Sentiment Analysis of Text

Our classifier is based on the ‘Wordnet-Affect’ database which is an extension of the ‘Wordnet’ natural language processing database. ‘Wordnet’ presents a linguistic resource for a lexical representation of affective knowledge by assigning 'A-Tags' which can extrapolate emotional states and expected responses (Strapparava et al. 2004).

4. Conclusion

Deployment in a fully functional game or a game in its beta testing stages will be required for testing our finished prototype. Through monitoring conversations between players and logging their emotions in real time through our system, developers can have a clear view of the emotional response being elicited from players and in turn know to alter the AI of the game accordingly. Whether a game’s intention is to train or entertain, every game intends to elicit some kind of emotional response from its players. Our middleware will provide useful player generated feedback to adapt a game’s intelligence in real time in order to move toward eliciting the desired emotion.

5. Bibliography

Park, Eunil., Ki Joon, Kim.,Shyam Sundar, S.,del Pobil, Angel P., (2012). Online Gaming with Robots vs. Computers as Allies vs. Opponents, ACM/IEEE international conference on Human-Robot Interaction, 205-206

Schuller, B.; Vlasenko, B.; Eyben, F.; Rigoll, G.; Wendemuth, A.; (2009), "Acoustic emotion recognition: A benchmark comparison of performances," Automatic Speech Recognition & Understanding, pp.552-557.

Strapparava, Carlo., Valitutti, Alessandro., (2004), "WordNet-Affect: an Affective Extension of WordNet", Proc. Intl. Conf. Language Resources and Evaluation, vol. 4, pp.1083 -1086 2004

Alarm Filtering through Similarity Measure

Avneesh Vyas [1], John Keeney [1], Enda Fallon[2]

[1] *LM Ericsson, Ireland*
{avneesh.vyas, john.keeney}@ericsson.com
[2] *Athlone Institute of Technology, Ireland*
efallon@ait.ie

1 Introduction

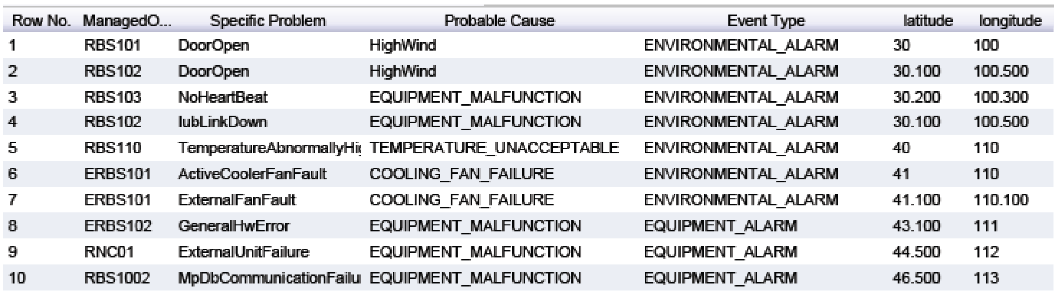
Currently, most commercial telecommunications network management systems require significant human intervention in managing the stream of alarms from the network. Fault Management (FM) systems receive alarms from the network and store them in internal database for analysis by the human operator. Most network operators make extensive use of statically-encoded human-expertise based systems to filter and correlated network faults into a form whereby they can be inspected and acted upon by either human-centric or automated management processes. While such an approach has worked to some extent until now, this cannot be expected to scale to already existing situations where millions of mobile devices and non-carrier-grade pico-cell stations need to be monitored and managed. Given the explosion of alarm rates, this manual approach of skimming through each alarm records is already inadequate. This results in an urgent need of new techniques and approaches that can reduce, if not completely eliminate, the human intervention for alarm management and thus improve the overall analysis efficiency of network operators. Machine learning techniques could be applied to assist in alarm filtering and correlation, and also to discover and rank best-practice corrective actions. While there exists a plethora of off-the-shelf machine learning tools and techniques that might assist in this problem, the main underpinning of any machine learning algorithm is the ability to calculate a similarity or distance metric between different cases (or, entities, inputs, outputs, clusters, patterns, etc.). For telecoms management, existing networks are much larger, more diverse, distributed and complicated and dynamic than most existing machine learning problem domains. Therefore, this simple-sounding problem of an adequate distance metric for comparing heterogeneous cases or entities is a fundamental blocker in the adoption of machine learning for telecoms management. In this work we present a preliminary approach to compare complex network alarms, with different types, schemas, and constituencies. Once the similarity of events can be determined, they can then be clustered into patterns, composed and correlated into more intelligible composite alarms, thus supported common processing and vastly assisting in root cause analysis and deduction of appropriate corrective actions.

2 Related Work

In an attempt to fix the problem of alarm floods, most existing Network Management solutions nowadays include some sort of an alarm/event correlation engine in their product offering. For example, Ericsson's OSS-RC [1] and IBM's Netcool [2] offer a similar optional alarm correlation component to their customers. To identify related alarms or events, these correlation systems refer to statically encoded expert rules authored by domain experts. The encoding of these rules is done using domain-specific rule languages and requires significant domain expertise and knowledge of individual network deployment characteristics. The major disadvantage of this approach is the overhead associated with maintenance of these expert rules. Additionally, these rules are deployment specific and brittle in terms of network situations they can handle. Consequently these rules offer no help even in slightly different scenarios, and degrade as the network characteristics evolve.

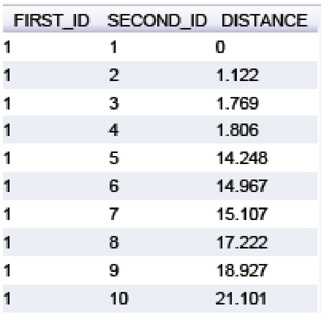
3 Solution and Results

An alternative way to correlate alarms is by generically comparing alarm context information (e.g. location, source, specific problem, time etc.). For instance, if a number of environmental alarms are generated from a common geographical area then there is a high probability that extreme weather may be the root cause. And a ‘temporary shut down’ of nodes from that area may be the most appropriate common action. Similarly, a number of alarms with a common ‘probable cause’ and ‘source object’ may require similar corrective procedure. So, we propose to employ similarity computation technique through which alarm vector be compared and their distance metric be computed. To perform this each alarm record must be modeled as a multi-dimensional vector of most-likely relevant alarm attributes. For this presentation, the alarm model is roughly based on a subset of 3GPP alarm information model and considering the mixed data types of comprising attributes, Mixed Euclidean Distances between alarms can be computed. Of particular note here is the use of sub-functions to calculate similarity of fields of similar type, where these sub-functions can be extended and to incorporate additional semantics about the fields (e.g. dynamic topology-awareness, locations-based reasoning, node-aware problem analysis etc.). Also, more accurate results can be achieved by tuning the weights of the individual field distances of the alarm model, where this weighting may itself be learned and evolved based on accuracy and uptake.



Row No.	ManagedO...	Specific Problem	Probable Cause	Event Type	latitude	longitude
1	RBS101	DoorOpen	HighWind	ENVIRONMENTAL_ALARM	30	100
2	RBS102	DoorOpen	HighWind	ENVIRONMENTAL_ALARM	30.100	100.500
3	RBS103	NoHeartBeat	EQUIPMENT_MALFUNCTION	ENVIRONMENTAL_ALARM	30.200	100.300
4	RBS102	IubLinkDown	EQUIPMENT_MALFUNCTION	ENVIRONMENTAL_ALARM	30.100	100.500
5	RBS110	TemperatureAbnormallyHi	TEMPERATURE_UNACCEPTABLE	ENVIRONMENTAL_ALARM	40	110
6	ERBS101	ActiveCoolerFanFault	COOLING_FAN_FAILURE	ENVIRONMENTAL_ALARM	41	110
7	ERBS101	ExternalFanFault	COOLING_FAN_FAILURE	ENVIRONMENTAL_ALARM	41.100	110.100
8	ERBS102	GeneralHwError	EQUIPMENT_MALFUNCTION	EQUIPMENT_ALARM	43.100	111
9	RNC01	ExternalUnitFailure	EQUIPMENT_MALFUNCTION	EQUIPMENT_ALARM	44.500	112
10	RBS1002	MpDbCommunicationFailu	EQUIPMENT_MALFUNCTION	EQUIPMENT_ALARM	46.500	113

Table 1 Alarm records on which similarity computation was performed



FIRST_ID	SECOND_ID	DISTANCE
1	1	0
1	2	1.122
1	3	1.769
1	4	1.806
1	5	14.248
1	6	14.967
1	7	15.107
1	8	17.222
1	9	18.927
1	10	21.101

Table 2 Vector distance between alarm 1 and other alarms

4 Conclusion

While this presentation focuses on finding a comprehensive and robust approach to calculate the similarity or distance between network alarms, such an approach will contribute significantly in wide variety of telecom network management use cases apart from alarm filtering and correlation. For example when alarms occur, such an approach can also be used to compare management actions to select from appropriate corrective actions applied when similar alarms occurred previously. Telecom management systems can also use this technique to predict future network trends with a quantifiable degree of certainty by comparing new unseen network situations with historical situations and trends.

References

[1] http://www.ericsson.com/ourportfolio/products/oss-rc, Ericsson OSS-RC

[2] http://www-142.ibm.com/software/products/us/en/ibmtivolinetcoolomnibus/, Tivoli Netcool/ OMNIbus

Development of a Reusable Learning Object for Use in Teaching Structural Engineering

Jason Corbett [1] & Paul Archbold [2]

[1] Athlone Institute of Technology
A00149143@student.ait.ie

[2] Athlone Institute of Technology
parchbold@ait.ie

Abstract

Reusable learning objects may be considered to be pre-developed digital learning activities that may be integrated into lessons, modules, and courses. Their advantages include their flexibility of content, their accessibility and the possibilities of using them as a means of evaluation. The aim of this project was to develop such an object for use in teaching structural engineering. A review of best practice in development of reusable learning objects was conducted, followed by a suitability analysis of various software packages available for use in the development of the resource.

A reusable learning object to be used in the teaching of the principle of second moment of area was developed. This interactive resource was made accessible through a virtual learning environment (Moodle) and accessed by students from engineering and non-engineering programmes in both AIT and GMIT, who provided feedback on the object via an online survey.

Keywords: reusable learning object, structural engineering, education, virtual learning environment

1 Reusable Learning Objects

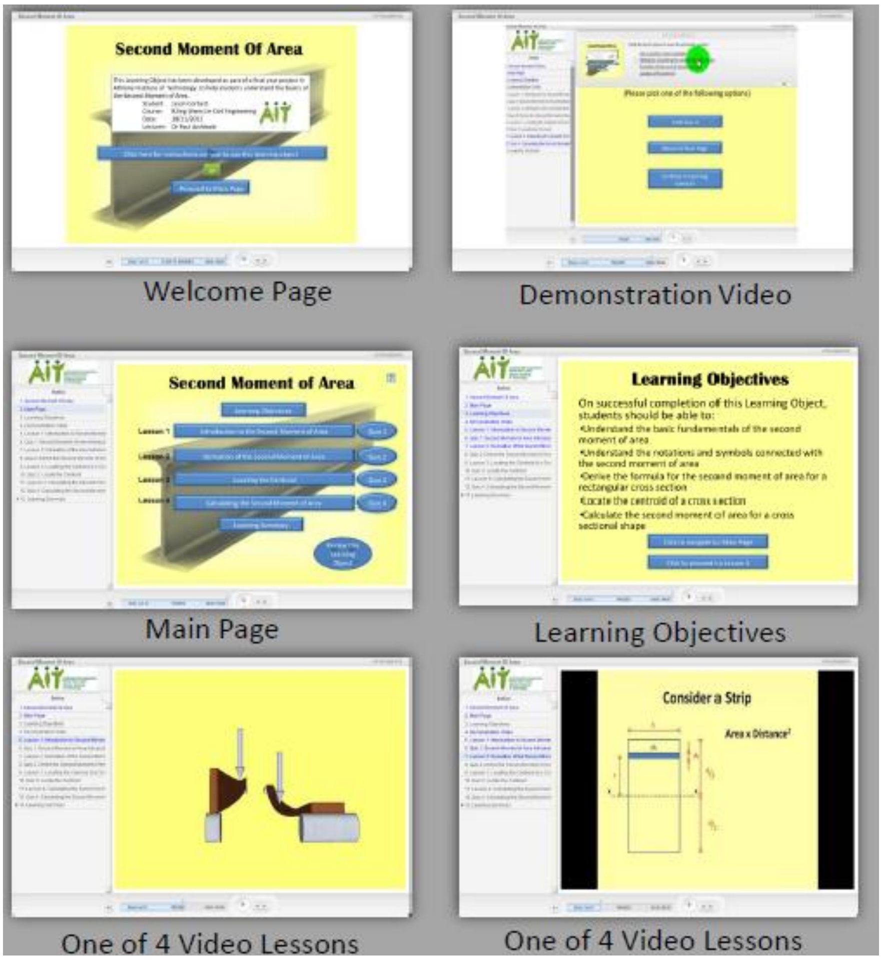
RLOs, as described by Billings (2010), are pre-developed digital learning activities that can be integrated into lessons, modules, and courses while Clyde (2004) stated that RLOs may be "chunks" of content, they may also be simulations, communication tools, assessment activities and learning management tools.

Currently, one of the ways in which faculty concerned about teaching attempt to improve their students' learning experience is through the use of technology in the classroom, with the most popularly visible instance of this being distance or e-learning. But less visible to the public eye and more challenging are individual faculty members' design and development of a plethora of exciting, unique, and innovative technology-based learning objects and methodologies that can have a significant effect on student learning outcomes. RLOs can be thought of as instruction-oriented software application modules that are either used in a stand-alone mode or embedded in larger applications such as online e-learning courses (Reisman 2009).

In recent years, the notion of RLOs or learning resources, have received significant attention in education communities. Initially inspired by object oriented programming practice in computer science, the idea appears to have materialised from traditional, instructional software design approaches issuing from professionals attempting to articulate more effective and economical strategies for management and reuse of resources in networked environments (Churchill 2007).

2 Development of Reusable Learning Object for use in Teaching Structural Engineering

A review of best practice in the development of reusable learning objects identified the CISCO Reusable Learning Object Strategy as an appropriate approach to the development of such resources and thus was adopted in this project. Further, suitability analysis of selected commercially available software packages to be used in the development of this resource considered parameters such as functionality, ease of use, developer’s expertise, visual impact, compatibility and cost. This led to the adoption of Google sketchup for developing the images and animation; Camtasia for recording and creating the tutorial video segments; Adobe Acobat was used for supporting documentation, while Articulate – an add-on to MS Powerpoint – was used as the host platform for all of the elements of the learning object. The learning object was then hosted in the Moodle Virtual Learning Environment. The object was successfully developed, reviewed and updated following user feedback. Figure 1 shows some screenshots from the finished reusable learning object.



Learning Objectives

On successful completion of this Learning Object, students should be able to:
•Understand the basic fundamentals of the second moment of area
•Understand the notations and symbols connected with the second moment of area
•Derive the formula for the second moment of area for a rectangular cross section
•Locate the centroid of a cross section
•Calculate the second moment of area for a cross sectional shape

Figure 1 Screenshots from Completed Reusable Learning Object

Conclusions

A reusable learning object was successfully developed for use in teaching the principle of second moment of area in structural engineering. This object utilised a number of commercially available software packages to create an interactive resource which was then made available to selected students via the virtual learning environment, Moodle. Feedback from students who used the resource was generally quite positive in terms of its

References

Billings, D.M. (2010) Using reusable learning objects. *Journal of Continuing Education in Nursing,* 41 (2) pp. 54-55.

Churchill, D. (2007) Towards a useful classification of learning objects. *Educational Technology Research & Development,* 55 (5) pp. 479-497.

Cisco (2003) Reusable Learning Object Strategy. *Designing and Developing Learning Objects for Multiple Learning Approaches.* [Online]. Available at: http://www.e-novalia.com/materiales/RLOW__07_03.pdf.

Clyde, L.A. (2004) digital learning objects. *Teacher Librarian,* 31 (4) pp. 55-57.

Evolving and Analyzing Strategies and Traits for the Hawk-Dove Game

Caroline Fennessy [1], Seamus Hill [2]

Department of IT, NUI, Galway, University Rd, Galway
[1] c.fennessy1@nuigalway.ie
[2] seamus.hill@nuigalway.ie

Abstract

This paper investigates the use of genetic algorithms to evolve strategies for the Hawk-Dove game theory model. The evolved strategies will be analyzed to identify various traits contained within the strategies. The paper will then examine the effects of these traits within strategies and to determine which individual traits or combination of traits, have the most impact. In addition this paper will look at the possibility of identifying and analyzing common traits that may exist in these optimal strategies.

Keywords: Hawk-Dove, Genetic Algorithms, Evolutionary Game Theory, Strategies, Traits

1 Introduction

The Hawk-Dove game is a traditional example of a game theory model used for studying strategic decision making among rational individuals [7]. In other game theory models players or strategies have competed against each other and the optimal strategies and common traits among them have been analyzed and documented. As a result this paper will focus on completing this work on the Hawk-Dove model to determine if superior strategies such as Tit for Tat and Pavlov will triumph as the main contenders yet again or will new or different strategies produce greater results.

In addition, it has been determined in other game theory models that common traits are present in optimal strategies, such as being that they can defend against defectors, and they can take advantage of the mutual cooperation with other superior players. This paper will test if optimal strategies contain these common traits in the Hawk-Dove game.

2 The Hawk-Dove Game

The Hawk-Dove model was first presented in Maynard Smith and Price, *The Logic of Animal Conflict* in 1973. The game was devised to model animal behavior where a contest over a sharable resource such as territory or food occurs [1]. Two players complete by playing one of two strategies, either hawk or dove. The strategy of the Hawk is said to be a fighter strategy in which aggression is first displayed, then he will fight until he wins or is injured. The strategy of the Dove is more peaceful; he will also display aggression at first but will retreat if the threat of escalation becomes real [3, 6].

As can be seen in Table 1, a payoff is associated with each contest.

	Hawk	Dove
Hawk	(V-C)/2	V
Dove	0	V/2

Table 1: Hawk-Dove Payoff Matrix

The resource is given the value V and the cost of losing a fight is C [3, 6]. If a Hawk meets a Dove, the Hawk wins and gets the resource. If a Hawk meets a Hawk, half the time he wins and half the time he loses, so the average outcome is the resource minus the cost of the fight divided by two. If a Dove meets a Dove they share the resource [3, 6]. Nash Equilibrium describes what move each player will make to maximize her score based on accurate assumptions about the other player's actions [5]. An outcome of a game is Pareto optimal if no one can be made better off without making at least one individual worse off. Strategies that are able to mutually cooperate are said to be Pareto optimal [5]. For a strategy to be evolutionarily stable, it must have no mutant strategy that is; an individual who adopts a novel strategy which can successfully invade [3].

3 Genetic Algorithm

A genetic algorithm is a search heuristic based on the mechanics of evolution. Each simulation begins with an initial population, where strategies are selected randomly. Each strategy represents a player in the game. In this case, a chromosome of 1's and 0's is used where a 1 represents defection/hawk and a 0 represents cooperation/dove. A fitness function, determines the success of these strategies. A form of *Natural Selection* is then performed using tournament selection whereby solutions with higher fitness scores have a greater probability of being selected. Tournament selection involves running several tournaments among a certain amount of individuals chosen at random from the population. The winner of each tournament i.e. the one with the highest fitness is selected for crossover. The genetic operator's crossover and mutation are then applied to form the next generation where randomly selected parents exchange parts of their strategies at a probability of 0.7 after which mutation will occur at a rate 0.001 to introduce diversity. The process is repeated until the specified number of generations has been reached. [4]

4 The Simulation

The simulations will consist of a population size of 20 individuals per generation. Each game will consist of 151 moves. Each experiment will consist of 50 generations. Forty experiments will be performed. [2, 5]

In this study a three-memory game is used. Each strategy can remember the three previous moves of both themselves and the other strategy; hence there are 64 possibilities for the three previous games. Since a strategy requires the history of the three previous moves, an extra 6 bits will contain hypothetical moves for the three previous games so as the strategy can decide how to make the initial move. [2, 5]

Each player or strategy will play in a round-robin tournament, in that each play all other strategies in the population but do not compete against themselves. Players or strategies will make simultaneous moves with points being awarded based on the outcome. After each game the points are calculated and the history of the three previous moves is updated. [2, 5]

After a tournament, players are ranked according to their fitness score and selection, crossover and mutation occur and a new population is generated to take part in the next tournament. [2, 5]

5 Results

The results of the 40 experiments will then be examined to review the difference between them and to observe what the optimal strategies are based on the highest average score and to observe if new or existing strategies are recognized.

Furthermore, analysis will be performed on the successful strategies to observe if they have traits in common; if they can defend against defectors, and if they can take advantage of mutual cooperation with other optimal strategies.

6 Conclusions

Dependent on results

References

[1] Maynard Smith, J. and Price, G.R. (1973). *The Logic of Animal Conflict*, Nature 246 (5427): 15-18.

[2] Axelrod, R. (1984). *The Evolution of Cooperation*, New York: Basic Books.

[3] Maynard Smith, J. (1982). *Evolution and the Theory of Games*, Cambridge: Cambridge University Press.

[4] Mitchell, M. (1998). *An Introduction to Genetic Algorithms*, Cambridge Massachusetts: MIT Press.

[5] Golbeck J. (2002). *Evolving Strategies for the Prisoner's Dilemma*, In Mastorakis N. E. (ed.), Advances in Intelligent Systems, Fuzzy Systems, and Evolutionary Computation pp. 299-306, WSEAS Press

[6] Johansson, S. J. (1999). *Game Theory and Agents*, Karlskrona, Sweden: Kaserntryckeriet AB

[7] Romp, G. (1997). *Game Theory Introduction and Applications*, New York: Oxford University Press

Layering Reality: Realistic Driving Simulation

Michael Brogan [1], Noel Daly [2], David Kaneswaran [3], Seán Commins [2],
Charles Markham [3], Catherine Deegan [1]

[1] Department of
Engineering,
Institute of Technology
Blanchardstown,
Dublin 15

[2] Department of
Psychology
NUI Maynooth,
Co. Kildare,
Ireland.

[3] Department of Computer
Science
NUI Maynooth,
Co. Kildare,
Ireland.

michael.brogan@live.ie

Abstract

All existing driving simulators are based on virtual worlds. This paper presents a driving simulator that uses images and navigation data acquired by a Mobile Mapping System to allow for psychological testing of participants using unaltered, augmented and redacted videos of road features.

Keywords: Real-world images, Driving simulation, Psychometric studies

1 Introduction

Applications of driving simulators are many, from driver training to medical and psychological evaluation [1]. However, all driving simulators that exist, from the most basic to the most advanced, are based upon virtual environments, such as are present in the area of video-game consoles. Despite massive advancements in the realism of these virtual environments, none are truly photorealistic [2]. This paper presents a driving simulator that uses image sequences as data input.

2 Driving Simulator and Simulator Interface

The simulator is constructed using a triple monitor set-up with a standard PC-based steering wheel and pedals. A Microsoft Windows 7-based PC runs the simulation software, allowing the user interface to be used to enter participant details and select the scenario video. The time between displaying sequential video frames is then linked directly to the pressure applied to the simulator's accelerator pedal.

3 Unaltered, Augmented and Redacted Image Sequences

A 14,000 frame video was manipulated to form three different scenarios: unaltered, differing speed limits and redacted road lines. Examples are shown in Fig. 1. Three stretches of road of similar geometry were chosen as sites for manipulation (unaltered, redacted and augmented). Each of the three locations was either (1) left unaltered, (2) augmented or (3) redacted; this was done to counter any bias in the road geometry that might serve to change driver behavior, rather than the experimental manipulation itself. All 30 participants, when driving, encountered each of these scenarios once only.

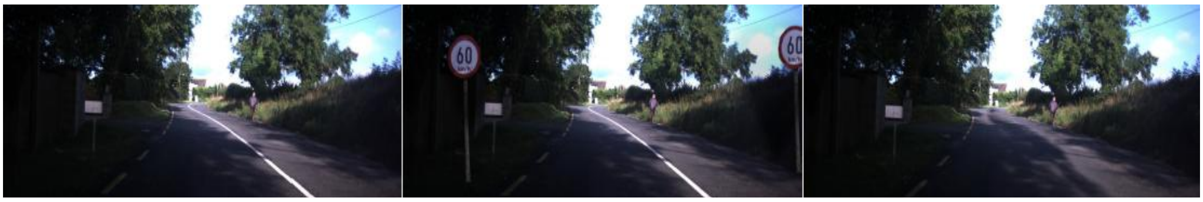


Figure 1: Unaltered, augmented (with inserted speed limit signs) and redacted (with erased road lines) images.

4 Preliminary Results

The current speed and current frame were recorded for each participant. Overall there were no observable differences between the redacted and augmented sequences, although the average speeds of the drivers corresponded to the geometry of the road.

The average driver speed across the test route is shown in Fig. 2. The aim of the experiment was to (a) observe if the participants would alter their speed based on the geometry of the road, and (b) react differently dependent on environmental factors, i.e. road speed limits and road markings.

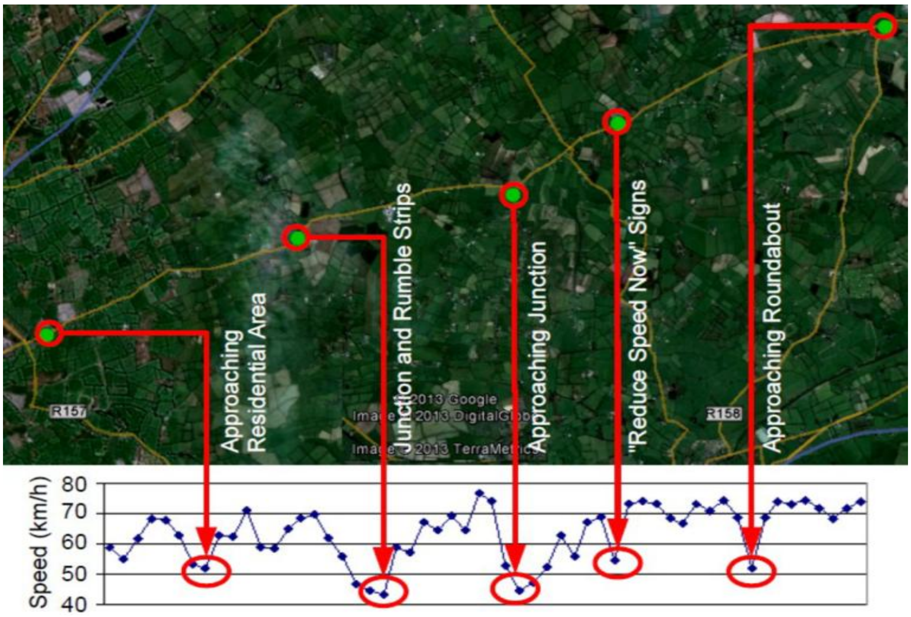


Figure 2: Changes in driver speed linked to road features and geometry.

Acknowledgements

The authors wish to acknowledge the Irish Research Council (IRC) for their continuing support.

References

[1] Tateyama, Y.; Mori, Y.; Yamamoto, K.; Ogi, T.; Nishimura, H.; Kitamura, N.; Yashiro, H.; "Car Driving Behaviour Observation Using an Immersive Car Driving Simulator," *P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2010 International Conference on*, vol., no., pp.411-414, 4-6 Nov. 2010.

[2] Murano, T., Yonekawa, T., Aga, M., and Nagiri, S., "Development of High-Performance Driving Simulator," *SAE Int. J. Passeng. Cars - Mech. Syst.* 2(1):661-669, 2009.

Low-Profile Mobile Mapping System

Michael Brogan, Seán Haughey, Simon McLoughlin, Catherine Deegan

School of Informatics and Engineering,
Institute of Technology Blanchardstown,
Dublin 15, Ireland.
michael.brogan@live.ie

Abstract

Most commercially available Mobile Mapping Systems require a dedicated vehicle and are of high-cost. They also require trained personnel to deal with the calibration, initialization and acquisition of data. The Mobile Mapping System described in this paper is vehicle independent, and is easy-to-use, calibrate and initialize. The system acquires full-color stereo images tagged with both Real-Time Kinematic Global Positioning and Inertial Measurement Unit data; data acquisition is possible at rates up to 10 Hz.

Keywords: Easy-to-use, Vehicle-independent, Mobile Mapping System, Short set-up time

1 Introduction

This paper describes a Mobile Mapping System (MMS) that acquires stereo image data alongside positional and orientation data at a rate of 10 Hz. The primary difference between this and existing MMSs is that this system is vehicle independent and has a short set-up time [1]. The system consists of two main subsystems; the stereo camera subsystem and the navigation subsystem, itself consisting of a Global Positioning System (GPS) antenna and receiver, and an Inertial Measurement Unit (IMU). The GPS allows the position of the system to be logged, and the IMU allows the orientation of the system to be logged. The GPS receiver is used to trigger the stereo camera, allowing for the acquisition of GPS and IMU tagged stereo images.

2 Data Acquisition System

The stereo camera subsystem used is a PGR Bumblebee XB3 stereo camera, consisting of stereo Bayer filter color camera sensors, each providing a maximum resolution of 1280x960. The XB3 uses time-locked pixels across the stereo images, allowing for absolute synchronization between the stereo image pairs, a feature of critical importance when acquiring stereo images from a moving vehicle, as a millisecond synchronization error at 80 km/h will cause a 2.2 cm positional error. The Bayer filter allows for a two-thirds reduction in raw image data, as full color 24-bit RGB images can be formed from the grayscale 8-bit images acquired directly by the camera [2]. This results in raw stereo image files of 2.34 MB, with an image data rate of over 23 MB per second. The camera is connected to the Global Positioning System (GPS) *via* a standard Hirose connection, and acquires a stereo image when a pulse generated by the GPS receiver is detected. Power and data transfer are achieved *via* the XB3's FireWire connection.

The navigation subsystem consists of two main parts; the GPS receiver and antenna, and the Inertial Measurement Unit (IMU). The GPS antenna and receiver are connected *via* a standard coaxial cable to form the positioning component of the navigation subsystem. Standard GPS receivers are capable of accuracies of around five meters, however, in a process known as Real Time Kinematic (RTK) correction streaming, this error can be reduced from five meters to less than three centimeters, typically around two centimeters. This allows the position of the system to be known and logged with an accuracy of two centimeters. The final part of the navigation subsystem is the IMU which allows for the orientation of the system to be recorded in three dimensions, i.e. system orientation in terms of Roll, Pitch and Yaw (RPY) angular deviation. The IMU is interfaced with the GPS receiver *via* the receiver's serial communications port. The GPS receiver and IMU are manufactured by NovAtel®, and are referred to collectively as a NovAtel® SPAN system [3]. This navigation system is interfaced to the control PC *via* a serial-to-USB connection. The serial-to-Hirose connection that interfaces the camera with the GPS receiver is connected to the receiver's serial I/O port. This port allows a Pulse Per Second (PPS) signal to be generated each time navigation data is logged. The receiver allows data logging at a number of predefined rates (e.g. 20 Hz, 10 Hz, 5 Hz, 2 Hz and 1 Hz). As the maximum frame rate of the XB3 camera is 16 Frames Per Second (FPS), and the closest maximum PPS rate is 10 Hz, the data acquisition system acquires image and navigation data at a rate of 10 Hz. A block diagram of the system is shown in Fig. 1.

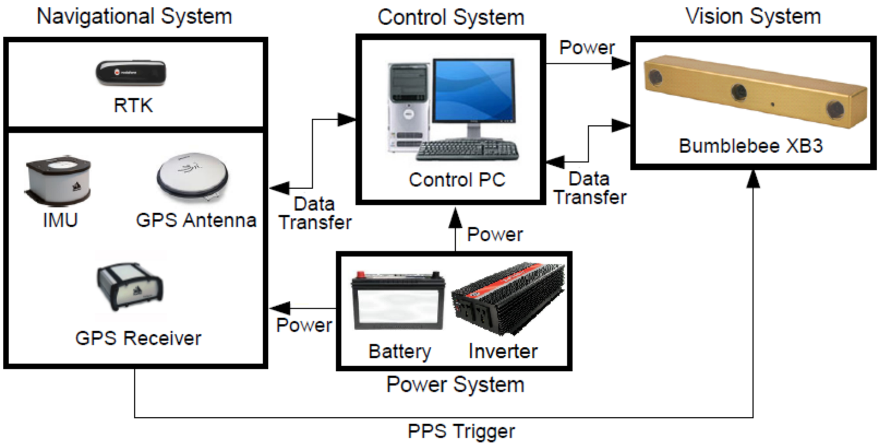


Figure 1: Block diagram of data acquisition system.

Once the camera, GPS antenna and IMU have been connected to the GPS receiver, and the receiver connected to the control PC, the system is ready to begin data acquisition. To begin acquisition, the GPS receiver is enabled; the RTK corrections and the IMU are initialized. The navigation data is then logged. The system used in the acquisition of this data is vehicle independent; all equipment is placed in the boot of the vehicle, and powered *via* an inverter through the car battery. The only observable components of the system are the GPS antenna, which sits on the inside rear shelf of the vehicle, and the camera system, which is mounted using industrial suction mounts on the bonnet of the vehicle.

3 Conclusions

This paper has presented a low-profile Mobile Mapping System that acquires full color stereo images alongside both position and orientation data at a rate of 10 Hz. In comparison to other MMSs, this system is fully portable, does not require a dedicated vehicle, has a short set-up and calibration time, and is user-friendly.

Acknowledgements

The authors wish to acknowledge the Irish Research Council (IRC) for their continuing support.

References

[1] C. Ellum and N. El-Sheimy, "Land-based mobile mapping systems", *Photogrammetric Engineering and Remote Sensing*, vol. 68, no. 1, pp. 13-28, 2002.

[2] Point Grey Research Bumblebee XB3 [Online: http://ww2.ptgrey.com/stereo-vision/bumblebee-xb3] [Last accessed: 2 April 2013]

[3] NovAtel SPAN System [Online: http://www.novatel.com/products/span-gnss-inertial-systems] [Last accessed: 2 April 2013]

Setting Up Automated Testing in an iOS Development Environment: An Experience Report

Dominic Maguire and Máté Rácz

Data Mining & Social Computing Research Unit, TSSG,
Waterford Institute of Technology.
dmaguire@tssg.org, mracz@tssg.org

Abstract

This poster outlines our experiences in setting up an automated testing infrastructure for an iOS development project. We outline the methods, tools and workarounds that were required to provide unit, application and acceptance testing in addition to code coverage reporting and application distribution.

Keywords: iOS Development, Automated Testing, Agile

1 Introduction

Martin Fowler defined continuous integration as "a software development practice where members of a team integrate their work frequently, usually each person integrates at least daily - leading to multiple integrations per day. Each integration is verified by an automated build (including test) to detect integration errors as quickly as possible" [1]. Automated testing is an integral part of the iterative continuous integration process and affects many of its practices. It achieves reliability at the object level via unit testing although the existence of unit tests does not guarantee reliability. Acceptance tests extend the scope of automated testing by creating a new paradigm, known as behaviour-driven development (BDD) [2]. In BDD, a customer's expectations of the application's behaviour forms the basis of acceptance tests which are then written and executed to confirm that the application does indeed meet its specifications.

1.1 The Problem

It has been noted that iOS development projects are not "first-in-class" when it comes to managing the quality of the software produced [3]. Furthermore, running iOS application tests from the command line is currently not officially supported by Apple [4] which presents a challenge in setting up a remote continuous integration server. In this poster we outline how we achieved a workable automated testing solution.

2 Our Approach

2.1 Unit & Application Testing

Apple have historically differentiated between Logic and Application tests. Logic tests are created for single methods whereas application tests involve the whole shebang, including UI elements. In order to integrate both types of test into the continuous integration process they need to be run from the command-line. The `RunPlatformUnitTests` script found at `/Developer/Platforms/-iPhoneSimulator.platform/Developer/Tools` shows that TEST_HOST is only set when running the tests through Xcode, not from the command line. If the host is not set, the tests do not run. Some advocate editing the script in situ to allow the tests to run [5]. We took the approach of Stuart Gleadow [6] in copying the above script into our project rather than hacking the system version. Two caveats are that the simulator needs to be closed when running the tests and the script needs to be executableon the build server. We achieved the latter via svn.

2.2 Acceptance Testing Using Frank and Cucumber

The combination of Frank and Cucumber seemed to be the most widely used solution for iOS acceptance testing. Frank is "Selenium for native iOS apps". It allows developers to write automated acceptance tests which verify the functionality of a native iOS app. Cucumber is a tool that executes plain-text functional descriptions as automated tests.

2.2.1 Setting Up Frank and Cucumber

In order for Frank to work, the accessibility features of the development/build machine need to be turned on. These are found at System Preferences -> Universal Access. We followed the instructions provided by Pete Hodgson [7]: (i) Install the frank-cucumber gem by running `sudo gem install frank-cucumber`, (ii) Run `frank setup` to create a Frank subdirectory which contains everything necessary to Frankify the app, (iii) Run `frank build` to create a Frankified version of the app, (iv) Run `frank launch` to launch the Frankified app in the simulator, (v) Check that a Frankified version of the app is running by using the command `frank inspect`. This opens up Symbiote in the web browser (`http://localhost:37265`). Symbiote is a web app embedded in the Frankified project that allows the inspection of the current state of the app as it is running.

3 Conclusion

This poster has shown our experience in setting up an automated testing infrastructure for iOS development. It cannot be guaranteed that the above steps will work with future versions of Xcode but we are currently able to provide a workable iOS continuous integration solution using the above methods.

References

[1] Martin Fowler. Continuous integration. 2006. http://www.martinfowler.com/articles/continuousIntegration.html.

[2] Jez Humble and David Farley. *Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation*. Addison-Wesley Professional, 1st edition, 2010.

[3] Cyril Picat. ios dev: How to setup quality metrics on your jenkins job? 2012. http://blog.octo.com/en/jenkins-quality-dashboard-ios-development/.

[4] Manuel Binna. Continuous integration of ios projects using jenkins, cocoapods, and kiwi. 2013. http://9elements.com/io/index.php/continuous-integration-of-ios-projects-using-jenkins-cocoapods-and-kiwi/.

[5] Gerard Condon. Gerard Condon's Blog. 2012. http://www.gerardcondon.com/blog/2012/09/20/further-jenkins-setup-code-signing/.

[6] Stuart Gleadow. Stuart Gleadow's Blog. 2012. http://www.stewgleadow.com/blog/2012/02/09/running-ocunit-and-kiwi-tests-on-the-command-line/.

[7] Pete Hodgson. Painless iOS Testing With Cucumber. 2012. http://testingwithfrank.com/.

Utilization of Text And Document Context in Large Content Repositories to Aid Environment Specific Search

Mr. Kyle Goslin [1], Dr. Markus Hofmann [2]

Institute of Technology Blanchardstown,
Blanchardstown Road North,
Dublin 15

kylegoslin@gmail.com [1]

markus.hofmann@itb.ie [2]

Abstract

Throughout the lifespan of any educational institution or business large numbers of documents are created and stored often in different document repositories. These document repositories can be as simple as a number of documents in a logical order or as complex as an interactive web based environment that allows users to upload files and append additional metadata and description text to the documents.

In early search, a focus has been placed upon this descriptive text as a means of identifying the document. As search progressed an additional focus was placed upon the body of the text as a more accurate means of identifying the relevance of the document to the user. With the inherent complexity of more recent repositories, during its lifespan a large quantity of layout, ordering and relationships of documents is created by its users. This data creates a natural context for the documents added to the repository.

This study outlines this utilization of this context of text in documents and also of documents to their native repository incorporating logical ordering and hierarchies in Moodle, a Virtual Learning Environment (VLE) to aid environment specific search.

Keywords: Document Repositories, Text Processing, Information Retrieval, Text Similarities

1 Introduction

With the growth in the number of document repositories that are being used on a daily basis the number of documents accessible to users during a search has also grown. This increase in documents provided the foundation problem of how the relevance of these documents to the user be correctly assessed. The concept of creating descriptive titles and appending additional metadata has been heavily researched [1, 2] and a number of different tools to automatically generate additional metadata for documents to aid search have been created [3].

This additional metadata can be key terms [4, 5] extracted from the documents, segments of text from defined areas of interest such as abstracts or through the classification [6] of the documents. All of these methods however are specifically content specific metrics to identify the relevance of documents. Although the collection of words and sentences in documents are of fundamental importance, additional measurements can be included to aid the description generation process of a document. This study outlines the process of extracting content from a wide variety of common document types to create an interconnected tree of document content with inherent document context extracted from the document repository. This approach provides an additional level of understanding of the documents in the repository to aid future searching.

2 Text Extraction Quality Assessment of Data

Content added to document repositories could be in any number of different common formats such as Microsoft Word, Microsoft PowerPoint and Adobe PDF. Each format however can be typically divided into two main categories, those of which as the content of documents as text based objects (plain text, XML or a binary representation) and retain the original content and those that represent the content as fully image based representations.

This poses the initial issue of extracting the content from each of these file types, to retain as much information as possible. A number of different tools can be used to extract this content from documents such as by identifying text objects in documents through an understanding of the document format or through the use of Optical Character Recognition (OCR) to convert image based representations of words into text. Combination approaches exist that utilize and understanding of the document format and apply OCR to images in files to yield the best results.

3 Text Features & Utilization of Context

The structures of course pages and system relationships extracted from the local database provide an additional level of understanding of ordering and relationships between documents. When extracted, this gathered context provides a rich base of additional knowledge for each document when viewed as an element of the repository. When document formatting and styles are applied an additional insight can be added to the text to further enrich the relevance of words in documents and documents to a specific search query. Identifying a reference point such as a course or individual document in the learning environment provides a starting point and natural narrowing of the searchable dataset.

4 Conclusion & Future Work

Once a document repository is assessed and document content and formatting is extracted along with inherent document structures an additional insight can be gained into the relevancy of the content to a specific query or relevance assessment made by a user. This study has lead to the quality assessment of text extraction tools to retain as much of the original text as possible with original formatting. Graphs outlining the relationships and context of data in document repositories have been created. A number of possible applications exist such as search enrichment, are currently being developed as part of this on going project.

References

[1] Pallickara, S.L. et al. (2010). Efficient Metadata Generation to Enable Interactive Data Discovery over Large-Scale Scientific Data Collections, *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, vol., no., pp.573-580, Nov. 30 2010-Dec. 3 2010.

[2] Edvardsen, L.F.H., et al. (2009). Using automatic metadata generation to reduce the knowledge and time requirements for making SCORM learning objects, *Digital Ecosystems and Technologies, 2009. DEST '09. 3rd IEEE International Conference on*, vol., no., pp.253-258, 1-3 June 2009.

[3] Paolo Bolettieri, et al. (2007). Automatic metadata extraction and indexing for reusing e-learning multimedia objects. In *Workshop on multimedia information retrieval on The many faces of multimedia semantics* (MS '07). ACM, New York, NY, USA, 21-28.

[4] Thomas Bohne, et al. (2011). Efficient keyword extraction for meaningful document perception. In *Proceedings of the 11th ACM symposium on Document engineering* (DocEng '11). ACM, New York, NY, USA, 185-194.

[5] Jiajia Feng, et al. (2011). Keyword extraction based on sequential pattern mining. In *Proceedings of the Third International Conference on Internet Multimedia Computing and Service* (ICIMCS '11). ACM, New York, NY, USA, 34-38.

[6] Han Lu, et al. (2008). The effects of domain knowledge relations on domain text classification, *Control Conference, 2008. CCC 2008. 27th Chinese*, vol., no., pp.460-463, 16-18 July 2008.

Evaluating a Driver's Road Accident Risk using National Map, Accident and Journey Data.

D. Kaneswaran[1], M. Brogan[2], C. Markham[1], S. Commins[1], C. Deegan[2]

[1] NUI Maynooth, Kildare, [2] IT Blanchardstown, Dublin 15
david.kaneswaran@nuim.ie

Abstract

This study is part of a collaborative research project that is investigating the role of road geometry and driver behavior in road accidents. The following paper brings together three large datasets in order to estimate accident occurrence rates on Irish roads. The combination of these three datasets within a single application produced a mapping system with the ability to select particular roads and accident information. This information includes accident risk variations between road types.

Keywords: Road Risk Analysis, Road Accidents, Road Journeys, Road Accident Rate, Road Data.

1 Introduction

Recording of traffic accident events is standard practice in most developed countries. In 2010 the Road Safety Authority (RSA) recorded over 5,780 accident events in Ireland. The RSA provides online access to a comprehensive set of accident statistics, including a display of a national geo-coded pin map. However, this dataset does not facilitate the selection of *individual* roads to determine the accident count per road. The first aim of the project was to develop an approach, by which road selection and accident count could be automated, in a standalone application. The second was to apply this approach to other national road data sources to provide a more accurate estimation of road accident risk.

2 Method

The RSA provided an accident dataset that included lighting conditions, severity of accident, car description etc. From this dataset the time and location information were extracted. A second set of data was acquired from Open Street Map **(**OSM) [2]. The OSM data were parsed to create a database describing the road network of Ireland. A visualization of this database was then developed in C#. By correlating the RSA and the OSM data on the display, it is possible to see accident events overlaying the Irish road network (See Figure 1).

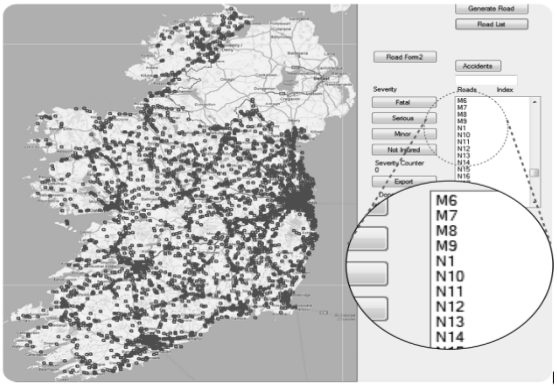


Figure 1. New mapping system for accident data

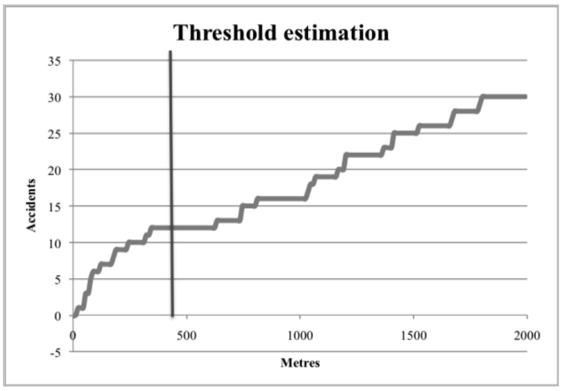


Figure 2. Thresholding of accidents on the M8

In order to associate an accident event with a specific road, an algorithm to calculate the threshold value was developed. This implemented an iterative approach to gradually increase the area associated with the road and any nearby accidents. As the area increased, a number of plateaus in the graph were observed, each plateau corresponding to the road and its nearest neighbor. The threshold value was then set near the center of the closest plateau. The number of accidents associated with the road, level out at this threshold; an accident value of 12. Figure 2 shows the threshold calculation for a selected road (M8). A threshold of approximately 478 meters was used to automate a distance limit for each accident event. This approach allowed an ‘accident per road’ calculation to be applied to other roads. The third dataset was acquired from the NRA’s Automatic Counter Statistics [3]. A selection of specific road information from this source was applied to the ‘accident per road’ calculations.

3 Results and Conclusion

The combination of the three datasets within a single application produced a mapping system with the ability to select particular roads and display location-specific accident information. Figure 3 shows an accident-to-traffic-rate calculation for 8 roads (4 M roads and 4 N roads). By applying this technique to specific roads it was possible to determine accident risk variations between road types. Figure 3 demonstrates that the selected N roads have a higher average accident rate (Mean: 0.56, Std. Dev.: 0.45) than the selected M roads (Mean: 0.11, Std. Dev.: 0.06). In addition, Figure 4 shows the hourly national accident rate (per annum) against the hourly traffic rates of an individual M road per annum. These findings show a linear correlation (R^2= 0.934) between accident rate and traffic levels. With no traffic, the risk is 0.218 accidents per hour (per day), per year. High traffic volumes increases the risk to 1.102 accidents per hour, per day, per year, this corresponds to a risk factor increase of 5.05. Other research has shown that congested traffic conditions may increase road accident risk in agreement with Figure 4, however this work also suggests the severity of the accident may differ under uncongested conditions [4].

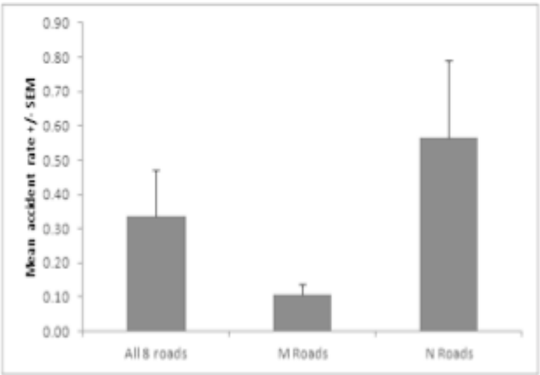


Figure 3. Mean accident rate per 1000 journeys per 100km

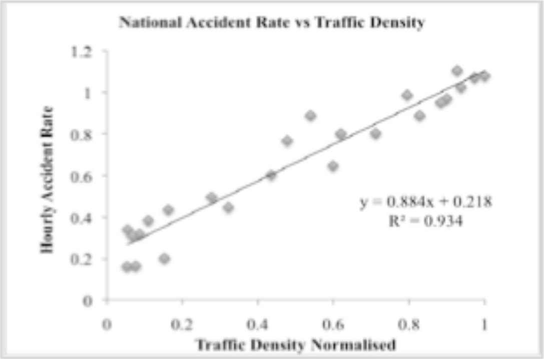


Figure 4. Estimated accident rate as a function of M8 hourly traffic rate

By combining three datasets it is possible to estimate both the road accident risk per individual road and a future accident rate prediction with respect to journey rates. These findings are supported by other research that shows the linear relationship between accident rate and traffic volume [5].

Acknowledgements: The authors wish to acknowledge the RSA research department for the provision of a full accident dataset and would also like to acknowledge the Irish Research Council for their support.

References

[1] Road Safety Authority (2012) “Statistics” [online] Available: http://www.rsa.ie/en/RSA/Road-Safety/Our-Research/ [accessed 1 November 2012].

[2] Open Street Map (2012) “WikiProject Ireland” [online] Available: http://wiki.openstreetmap.org/wiki/WikiProject_Ireland [accessed 15 November 2012].

[3] National Roads Authority (2012) “Automatic Counter Statistics” [online] Available: http://nraextra.nra.ie/CurrentTrafficCounterData/index.html [accessed 1 November 2012].

[4] Noland, Robert B. and Quddus A, Mohammed. (2005). Congestion and Safety: A Spatial Analysis of London. Transport research Part A: Policy and Practice: Volume 39, Issues 7–9, 737-754

[5] PIARC (2007) Technical Committee on Road Safety: PIARC Road Accident Investigation Guidelines for Road Engineers Manual.