Computing and Information Systems

2ND ANNUAL DOCTORAL COLLOQUIUM

Program and Proceedings

23rd July 2014
Welcome to the CIS 2014 Doctoral Colloquium!

Welcome to the 2nd Computing and Information Systems (CIS) Doctoral Colloquium (DC) at the University of Melbourne. CIS DC offers PhD students a tailored full-day forum where they are given the opportunity to present and discuss their research.

To our delight, we have experienced a 48% increase in the number of submissions this year. This extraordinary participation encouraged us to put our best foot forward in organising the event, so that all attendees would benefit highly from it.

This year, the CIS DC has crafted a stimulating environment for students to broaden their network and receive supportive and constructive feedback on their research plans. We have achieved this by advocating for extensive involvement from University industry partners along with internal and external senior researchers at the CIS department. The CIS DC 214 offers a diverse academic menu comprising of paper presentations, interactive posters, and flash talks, headlined by keynote speaker, and capped off by conferring awards to the best five paper and poster presentations.

CIS DIC 2014 was organized by a group of RHD students at the CIS department of Melbourne University. As chair of the Organising Committee, it is my privilege to thank the team for the countless hours of effort they have put in over the course of the past 5 months.

Also, a special thank you to Professor Justin Zobel and Rhonda Smithies for their continuous support.

Once again, welcome to CIS DC 2014!

We all look forward to interacting with you today!

Behnaz R. Yeganeh, Committee Chair
Conference Location: Alan Gilbert Building

Building Number: 104
Street Address: Grattan Street
Campus Map reference: M13

Map of Conference Venue

Image source: en.wikimedia.org
Leaflet | Data, imagery and map information provided by MapQuest, OpenStreetMap and contributors.
### Conference Program

**Morning session**

- **8:30** | Registration and Arrival Tea/Coffee | Foyer, Level 1
- **9:00** | Opening and Welcome Address
  Prof. Justin Zobel | Lecture Theatre 2
- **9:15** | Keynote: Connecting theory & practice - a lesson from bio-informatics
  Dr. Tom Conway | Lecture Theatre 2

#### Cloud Computing (Session 1A)
- **10:00** | A Performance Optimization Scheme for Mobile Cloud
  Bowen Zhou | Lecture Theatre 3
- **10:15** | Resource Overbooking in IaaS Cloud Provider
  Chanh Nguyen |
- **10:30** | Fault-Tolerant Workflow Scheduling Using Spot Instances on Clouds
  Deepak Poola, Kotagiri Ramamohanarao and Rajkumar Buyya |
- **10:45** | Internet of Things: A Roadmap for Future Internet Direction
  Farzad Khodadadi |
- **11:00** | Energy Efficient Software Defined Clouds
  Jungmin Son |
- **11:15** | Resource Provisioning and Scheduling Algorithm for Workflows on Clouds
  Maria A Rodriguez and Rajkumar Buyya |
- **11:30** | Virtual Machine Customization and Task Mapping Model for Efficient Allocation of Cloud Data Center Resources
  S. F. Piraghaj, R. N. Calheiros and R. Buyya |
- **11:45** | Integrated Framework for Cloud-based Interoperability between Social Network Websites
  Safiollah Heidari and Rajkumar Buyya |
- **12:00** | SLA Based Resource Scheduling for Big Data Applications
  Yali Zhao |
- **12:15** | Virtual Machine Allocation Policies against Co-resident Attacks in Cloud Computing
  Yi Han, Jeffrey Chan, Tansu Alpcan and Christopher Leckie |

#### Information Systems (Session 1B)
- **10:00** | The influence of personality traits on online information behaviour during political campaigns
  Miguel Wood, Simon Milton and Graeme Shanks |
- **10:15** | A conceptual model for analysing informal learning in online social networks for health professionals
  Xin Li and Kathleen Gray |
- **10:30** | Task Complexity and Codification
  Marion Zalk, Rachelle Bosua and Rajeev Sharma |
- **10:45** | Online Medical Consultation: A review of current practices
  Ibrahim AL-Mahdi |
- **11:00** | Performance evaluation of spectral unmixing method based on different ways of training data selection
  Liyan Liu, Andrey Kan and Christopher Leckie |
- **11:15** | Retinal Cross Sectional Layer Segmentation using Optical Coherence Tomography
  Md Akter Hussain, Alauddin Bhuiyan and Kotagiri Ramamohanarao |
- **11:30** | An Effective Automated System for Grading Severity of Retinal Arteriovenous Nicking in Colour Retinal Images
  Pallab Kanti Roy, Uyen T. V. Nguyen, Alauddin Bhuiyan and Kotagiri Ramamohanarao |
- **11:45** | Shooting the Breeze Involves neither Shooting nor a Breeze: Predicting the Compositionality of Multiword Expressions
  Bahar Salehi |
- **12:00** | Semi-supervised Multilingual POS tagging
  Long Duong |
- **12:15** | One-Way Games
  Andres Abeliuk, Gerardo Berbeglia and Pascal Van Hentenryck |
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
<th>Location</th>
<th>Speaker(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>12:30</td>
<td><strong>Lunch Break</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:15</td>
<td>Developing an opportunistic online psychosocial screening tool for young people in general practice</td>
<td>Lecture Theatre 3</td>
<td>Marianne Webb, Greg Wadley and Sylvia Kauer, Lena Sanci</td>
</tr>
<tr>
<td>13:30</td>
<td>Help me keep my stuff safe</td>
<td>Lecture Theatre 3</td>
<td>Sarah Ellen Webber</td>
</tr>
<tr>
<td>13:45</td>
<td>Understanding the user experience of running with mixed reality stories</td>
<td>Lecture Theatre 3</td>
<td>Alexander Kan</td>
</tr>
<tr>
<td>14:00</td>
<td>Non-verbal Interactions in Domestic Video Conferencing</td>
<td>Lecture Theatre 3</td>
<td>Behnaz R. Yeganeh</td>
</tr>
<tr>
<td>14:15</td>
<td>Won’t you be my neighbour? Clustered borrowing in libraries as evidence for shelf browsing</td>
<td>Lecture Theatre 3</td>
<td>Dana McKay</td>
</tr>
<tr>
<td>14:30</td>
<td>Supporting Tele-consultation through Tangible Interfaces</td>
<td>Lecture Theatre 3</td>
<td>Deepti Aggarwal</td>
</tr>
<tr>
<td>14:45</td>
<td>Experiences with Mobile Mental Health Therapies</td>
<td>Lecture Theatre 3</td>
<td>Fernando Estrada</td>
</tr>
<tr>
<td></td>
<td><strong>Database, Machine Learning &amp; Security (Session 2B)</strong></td>
<td>Lecture Theatre 2</td>
<td>Improved k-centre Clustering with MapReduce</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Jessica McClintock</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Designing Organization-Aware Agents using OJAzzIC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Kathleen Keogh</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Reliable power transmission networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Rodrigo R. Gumucio E., Carleton Coffrin and Pascal Van Hentenryck</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Neighborhood Component Analysis for Feature Learning</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Zay Maung Maung Aye, Kotagiri Ramamohanarao and Benjamin Rubinstein</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mining User Interactions and Activities using Mobile Data while Maintaining Privacy</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Anthony Quattrone</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Privacy-Preserving Collaborative Anomaly Detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sarah M. Erfani, Yee Wei Law, Shanika Karunasekera, Christopher A. Leckie and Marimuthu Palaniswami</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Automatically Recognizing Places of Interests/Activities from Unreliable GPS Data and Smart Phone Sensors</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tanusri Bhattacharya, Lars Kulik and James Bailey</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td>Afternoon Tea &amp; Coffee</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Human Computer Interaction (Session 3A)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Database, Machine Learning &amp; Security (Session 3B)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lecture Theatre 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lecture Theatre 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td>Communication Technologies during Family Mealtime</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hasan Shahid Ferdous</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:45</td>
<td>Screen Ecologies and the Future of Domestic Gaming</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Marcus Carter, Bjorn Nansen and Martin Gibbs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:45</td>
<td>Principled Dictionary Pruning for Low-Memory Corpus Compression</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jiancong Tong</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:45</td>
<td>Predicting at-risk students in Massive Open Online Courses</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Jiazhen He, James Bailey, Rui Zhang and Benjamin Rubinstein</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Software Engineering (Session 3C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lecture Theatre 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Loop Untangling</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kathryn Francis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:15</td>
<td>Automated debugging using program spectra</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Neelofar, Lee Naish and Rao Kotagiri</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>Review of Interdependence in Coactive Design</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ronal Singh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:45</td>
<td>Improved Feature Transformations for Classification using Density Estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yamuna Kankanige and James Bailey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:45</td>
<td>FILTA : Better View Discovery from Collections of Clusterings via Filtering</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yang Lei, Nguyen Xuan Vinh, Jeffrey Chan and James Bailey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:45</td>
<td>Bandwidth-Efficient Convergence of Gossip-based Recommendation Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Irum Fahim Bukhari, Aaron Harwood and Shanika Karunasekera</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:45</td>
<td>Refreshments</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:45</td>
<td>Flash Talks</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lecture Theatre 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:45</td>
<td>Poster Exhibition</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Foyer, Level 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:30</td>
<td>Closing and Award Ceremony</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Evening session

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Lecture Theatre 3</td>
</tr>
<tr>
<td></td>
<td>Lecture Theatre 3</td>
</tr>
<tr>
<td></td>
<td>Foyer, Level 1</td>
</tr>
<tr>
<td></td>
<td>Closing and Award Ceremony</td>
</tr>
</tbody>
</table>
What do you think about that?

Tweet us @cis_dc and share your thoughts on the talk / presentation / poster you are listening to!

The hashtag to use is #cisdc2014
Connecting theory & practice - a lesson from bio-informatics

A divide often exists between the theoreticians and the “practical” researchers in a Computer Science department, with the former stereotypically seeing their work as real computer science, and the work of their colleagues as mere programming. The converse view is that the work of the theoreticians is irrelevant compared to the useful work of their colleagues. This talk will show by example how the work of the theoreticians is vital to the production of useful methodologies for real problems. The concrete example is the problem of de novo assembly – a core problem in bioinformatics. Along the way we will discuss a bit of biology, a bit of information theory, a bit about computer architectures, and correctness.

Bio Sketch
Dr. Tom Conway undertook his doctoral research in the area of Logic Programming Language implementation, specifically looking at models for automatic parallelization of programs, semantics of threads and profiling. After completing his PhD, he worked for Multimedia Database Systems – a spin-off from RMIT University developing advanced text search techniques. After leaving MDS, he then took up a bioinformatics position at NICTA, delving in to advanced data structures and algorithms and applying them to bioinformatic analysis problems. Tom is currently leading the Genomics team at IBM Research Australia.
Poster Submissions

Analysis of Road Traffic Using Contrast Mining and GPS Trajectories
Xiaoting Wang, Christopher Leckie, Tharshan Vaithianathan

Answering Complex Questions in Neuroimaging Informatics
Aref Eshghishargh, Simon Milton, Andrew Lonie, Gary Egan

Audience Experience in Domestic Videogaming
John Downs

Compression and Visualization of Large Scale Graphs
Lida Rashidi, Christopher Leckie, Sutharshan Rajasegaran

Data Quality and Data Cleansing in Bioinformatic Databases
Qingyu Chen, Justin Zobel, Karin Verspoor

Energy and Carbon-Efficient Resource Management for Geo-Distributed Cloud Data Centers
Atefeh Khosravi and Rajkumar Buyya

Fragment Based Planning using Column Generation
Toby Davies, Adrian Pearce, Peter Stuckey, Harald Søndergaard

Information Security Strategy in Organisations
Craig Horne

Intelligence Driven Information Security Risk Management
Jeb Webb, Atif Ahmad, Sean Maynard, Graeme Shanks

Lazy and Eager Approaches for the Set Cover Problem
Lim Ching Lih, Alistair Moffat, Tony Wirth

Memory Efficient Local Outlier Detection in Data Streams
Mahsa Salehi, Christopher Leckie, Tharshan Vaithianathan

Preserving Location Privacy in Location-based Services - Obfuscation through Locality Preserving Hashing
Maryam Fanaeepour, Lars Kulik, Egemen Tanin

Privacy Aware Dynamic Ride Sharing
Preeti Goel, Lars Kulik, Kotagiri Ramamohanarao

Read Sequence Alignment Compression
Rodrigo Canovas, Alistair Moffat, Andrew Turpin

Secure Repair in Large Scale Distributed Storage Systems
Lakshmi J Mohan, Udaya Parampalli, Aaron Harwood

Spatio-Temporal Trajectory Simplification for Inferring Travel Paths
Hengfeng Li, Lars Kulik, Kotagiri Ramamohanarao

Transfer learning of a temporal bone performance model via anatomical feature registration
Yun Zhou, Ioanna Ioannou, Sudanthi Wijewickrema, James Bailey, Patorn Piromchai, Stephen O’Leary, Gregor Kennedy
# Paper Submissions

## Cloud Computing

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Performance Optimization Scheme for Mobile Cloud</td>
<td>14</td>
</tr>
<tr>
<td>Bowen Zhou</td>
<td></td>
</tr>
<tr>
<td>Resource Overbooking in IaaS Cloud Provider</td>
<td>15</td>
</tr>
<tr>
<td>Chanh Nguyen</td>
<td></td>
</tr>
<tr>
<td>Fault-Tolerant Workflow Scheduling Using Spot Instances on Clouds</td>
<td>16</td>
</tr>
<tr>
<td>Deepak Poola, Kotagiri Ramamohanarao and Rajkumar Buyya</td>
<td></td>
</tr>
<tr>
<td>Internet of Things: A Roadmap for Future Internet Direction</td>
<td>17</td>
</tr>
<tr>
<td>Farzad Khodadadi</td>
<td></td>
</tr>
<tr>
<td>Energy Efficient Software Defined Clouds</td>
<td>18</td>
</tr>
<tr>
<td>Jungmin Son</td>
<td></td>
</tr>
<tr>
<td>Resource Provisioning and Scheduling Algorithm for Workflows on Clouds</td>
<td>19</td>
</tr>
<tr>
<td>Maria A Rodriguez and Rajkumar Buyya</td>
<td></td>
</tr>
<tr>
<td>Virtual Machine Customization and Task Mapping Model for Efficient Allocation of Cloud Data Center Resources</td>
<td>20</td>
</tr>
<tr>
<td>S. F. Piraghaj, R. N. Calheiros and R. Buyya</td>
<td></td>
</tr>
<tr>
<td>Integrated Framework for Cloud-based Interoperability between Social Network Websites</td>
<td>21</td>
</tr>
<tr>
<td>Safiollah Heidari and Rajkumar Buyya</td>
<td></td>
</tr>
<tr>
<td>SLA Based Resource Scheduling for Big Data Applications</td>
<td>22</td>
</tr>
<tr>
<td>Yali Zhao</td>
<td></td>
</tr>
<tr>
<td>Virtual Machine Allocation Policies against Co-resident Attacks in Cloud Computing</td>
<td>23</td>
</tr>
<tr>
<td>Yi Han, Jeffrey Chan, Tansu Alpcan and Christopher Leckie</td>
<td></td>
</tr>
</tbody>
</table>

## Information Systems

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The influence of personality traits on online information behaviour during political campaigns</td>
<td>24</td>
</tr>
<tr>
<td>Miguel Wood, Simon Milton and Graeme Shanks</td>
<td></td>
</tr>
<tr>
<td>A conceptual model for analysing informal learning in online social networks for health professionals</td>
<td>25</td>
</tr>
<tr>
<td>Xin Li and Kathleen Gray</td>
<td></td>
</tr>
<tr>
<td>Task Complexity and Codification</td>
<td>26</td>
</tr>
<tr>
<td>Marion Zalk, Rachelle Bosua and Rajeev Sharma</td>
<td></td>
</tr>
<tr>
<td>Online Medical Consultation: A review of current practices</td>
<td>27</td>
</tr>
<tr>
<td>Ibrahim AL-Mahdi</td>
<td></td>
</tr>
</tbody>
</table>
# Paper Submissions

## Life Sciences

**Performance evaluation of spectral unmixing method based on different ways of training data selection**  
Liyan Liu, Andrey Kan and Christopher Leckie  
Page 28

**Retinal Cross Sectional Layer Segmentation using Optical Coherence Tomography**  
Md Akter Hussain, Alauddin Bhuiyan and Kotagiri Ramamohanarao  
Page 29

**An Effective Automated System for Grading Severity of Retinal Arteriovenous Nicking in Colour Retinal Images**  
Pallab Kanti Roy, Uyen T. V. Nguyen, Alauddin Bhuiyan and Kotagiri Ramamohanarao  
Page 30

## Database, Machine Learning & Security

**Shooting the Breeze Involves neither Shooting nor a Breeze: Predicting the Compositionality of Multiword Expressions**  
Bahar Salehi  
Page 32

**Semi-supervised Multilingual POS tagging**  
Long Duong  
Page 33

**One-Way Games**  
Andres Abeliuk, Gerardo Berbeglia and Pascal Van Hentenryck  
Page 34

**Improved k–centre Clustering with MapReduce**  
Jessica McClintock  
Page 35

**Designing Organization-Aware Agents using OJAzzIC**  
Kathleen Keogh  
Page 36

**Reliable power transmission networks**  
Rodrigo R. Gumucio E., Carleton Coffrin and Pascal Van Hentenryck  
Page 37

**Neighborhood Component Analysis for Feature Learning**  
Zay Maung Maung Aye, Kotagiri Ramamohanarao and Benjamin Rubinstein  
Page 38

**Mining User Interactions and Activities using Mobile Datawhile Maintaining Privacy**  
Anthony Quattrone  
Page 39

**Privacy-Preserving Collaborative Anomaly Detection**  
Sarah M. Erfani, Yee Wei Law, Shanika Karunasekera, Christopher A. Leckie and Marimuthu Palaniswami  
Page 40

**Automatically Recognizing Places of Interests/Activities from Unreliable GPS Data and Smart Phone Sensors**  
Tanusri Bhattacharya, Lars Kulik and James Bailey  
Page 41
## Paper Submissions

### Principled Dictionary Pruning for Low-Memory Corpus Compression

Jiancong Tong

### Predicting at-risk students in Massive Open Online Courses

Jiazhen He, James Bailey, Rui Zhang and Benjamin Rubinstein

### Improved Feature Transformations for Classification using Density Estimation

Yamuna Kankanige and James Bailey

### FILTA : Better View Discovery from Collections of Clusterings via Filtering

Yang Lei, Nguyen Xuan Vinh, Jeffrey Chan and James Bailey

### Bandwidth-Efficient Convergence of Gossip-based Recommendation Systems

Irum Fahim Bukhari, Aaron Harwood and Shanika Karunasekera

### Human Computer Interaction

#### Developing an opportunistic online psychosocial screening tool for young people in general practice

Marianne Webb, Greg Wadley and Sylvia Kauer, Lena Sanci

#### Help me keep my stuff safe

Sarah Ellen Webber

#### Understanding the user experience of running with mixed reality stories

Alexander Kan

#### Non-verbal Interactions in Domestic Video Conferencing

Behnaz R. Yeganeh

#### Won’t you be my neighbour? Clustered borrowing in libraries as evidence for shelf browsing

Dana McKay

#### Supporting Tele-consultation through Tangible Interfaces

Deepti Aggarwal

#### Experiences with Mobile Mental Health Therapies

Fernando Estrada

#### Communication Technologies during Family Mealtime

Hasan Shahid Ferdous

#### Screen Ecologies and the Future of Domestic Gaming

Marcus Carter, Bjorn Nansen and Martin Gibbs
## Paper Submissions

### Software Engineering

<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop Untangling</td>
<td>58</td>
</tr>
<tr>
<td>Kathryn Francis</td>
<td></td>
</tr>
<tr>
<td>Automated debugging using program spectra</td>
<td>59</td>
</tr>
<tr>
<td>Neelofar, Lee Naish and Rao Kotagiri</td>
<td></td>
</tr>
<tr>
<td>Review of Interdependence in Coactive Design</td>
<td>60</td>
</tr>
<tr>
<td>Ronal Singh</td>
<td></td>
</tr>
</tbody>
</table>
ABSTRACT
Computation outsourcing is one of the Cloud-based mobile augmentation approaches that offload the resource intensive computation to the cloud to enhance mobile devices computing capabilities. In this paper, we propose a performance optimization partitioning scheme for mobile cloud computation outsourcing.

Categories and Subject Descriptors
C.2.4 [Distributed Systems]: Client/server

Keywords
Mobile cloud computing, cloud computing, computation offloading, resource outsourcing, remote execution

1. INTRODUCTION
Mobile devices come with compromised hardware such as low processor capabilities, small memory capacity and finite battery life. These limits have hindered mobile devices capabilities for real-time, computation intensive applications such as voice recognition. To overcome this inherent problem, computation offloading to rich computing resources is introduced. Cloud computing, a new computing paradigm emerging in the recent years, has the potential to provide such resources.

1.1 Cloud computing
With the trend of more powerful and cheaper computing resources, a new computing model call cloud computing emerged leveraging virtualization technology to provide computing resource as a utility [3]. The services of cloud computing are divided into three categories. First, Infrastracter as a Service provides end users with the control over operating systems, storage, networks and other fundamental computing resources through virtualization. Platform as a Service provides programming APIs and environments to the developers to build applications while it manages storage, operating system, servers as well as networks. Software as a Service provides ready-to-use applications to end users.

1.2 Mobile cloud computing
Mobile cloud computing (MCC) is arising as a prominent research area that focuses on delegate or offload data and processing to surrogates in order to enhance the mobile applications as well as user experience [2]. There exist definitions of MCC based on different research perspectives. Here we refer to the definition that integrating mobile devices with cloud-based resources.

1.3 Performance Optimization Scheme
Static partitioning proposed by previous works has the drawback that since it is difficult to predict the execution environment, partitioning the program at the design time may not be optimal and cannot adapt performance in another execution environment. On the other hand, dynamic partitioning can impose large overhead on mobile devices due to continuously partitioning [1]. Also, computation offloading needs the cost analysis to support the system in making decisions on when and how to partition the program to where.

To tackle the issues mentioned above, we propose a performance optimization partitioning scheme and a cost analysis model to support the decision making. The scheme uses a partitioning algorithm that combines static partitioning analysis with dynamical profiling, and a optimization model. The model aims at making the best trade-off between energy consuming and transmission cost to finalize the partitions based on the execution data obtained by the dynamic profiling. The developer’s effort is involved to specify the potential methods as partitions being offloaded in prior to the execution of the program. At runtime, the dynamic profiler will monitor hardware, network and the execution time of each offloaded tasks and update the cost analysis model. Due to the limited resources on mobile devices, this is implemented on the remote cloud. When the program is executed to the nominated partitions, we use the cost analysis model to evaluate whether to offload the partition or running it locally. The cost analysis model can be achieved by profiling the previous executed tasks in terms of the execution time, energy consumption, required computing capacity, etc.

2. REFERENCES
Resource Overbooking in IaaS Cloud Provider

[Extended Abstract]

Chanh Nguyen
Department of Computing and Information Systems
University of Melbourne
lnguyen2@student.unimelb.edu.au

ABSTRACT

Infrastructure-as-a-service Cloud provider offers different purchasing models along with various service guarantees respectively to give customer the flexibility to optimize their costs. As a consequence, the diversification of purchasing options introduces a far more trivial problem for providers in which they need to decide how many available resource should be segmented into each pricing model so that the expected revenue is maximized. In this work, We address problem of maximized revenue through developing the optimal algorithms in which leveraging information from resource demand prediction to make efficiently decision on resource segmentation respecting to available resource on data center and the SLA promised to the customers so that the final objective is maximized revenue.

Categories and Subject Descriptors
H.4 [Distribution Systems]: Resource Management

Keywords
Resource segmentation, Cloud computing, Resource utilization, Hybrid prediction model

1. INTRODUCTION

To give customers flexibility to optimize their cost, Cloud providers offer resource instance under multiple pricing models with different level of service guarantees. For example, customers are able to flexibly select their wishing virtual instances from Amazon EC2 under one of pricing schemes such as reservation, on demand, and spot instances [1]. According to the on demand plan, users only need to pay for actual resource usage without any contractual long term commitments. This price is usually set statically or infrequently updated by cloud provider. In the reservation plan, users pay an upfront reservation fee to reserve resources for a specific period (e.g., one year, three years, etc.). In return, they receive a significant discount on the hourly resource usage price. Whereas the spot market plan is at the opposite side, allows users to submit periodically their bid to provider describing the price they are willing to pay for an amount of resource. The users gain access to their resource request as long as their bid exceeds the providers spot price which is also described periodically by provider. In general, on demand instances can generate highest revenue per resource however they suffer from future demand uncertainty, while reserved instances can provide a risk free upfront income and foster long term commitments by customers. However, a disadvantage of the reservation instances is that the provider is responsible to provide guaranteed availability for the reserved requests while customers might not utilize their reserved resource fully in during the reservation period. A related study of parallel computing workloads [2] evidence this fact by showing that more than half of all jobs use less than 20% of users requested capacity.

At the cloud provider perspective, there exists an opportunity for them to maximize their profits by leveraging the underutilized capacity available to instance requests originating from other pricing plans (on demand, spot). Therefore, the diversification pricing plans not only allow providers to efficiently attract customers from distinct goals and preferences in using their resource, but also open a new direction to gain more revenue. More concretely, cloud providers need to solve the non-trivial problem of demand segmentation optimization through the aforementioned different pricing schemes. Besides that, with the unused capacity available from reservation contracts, providers also need to find out the reasonable overbooking factor so that they efficiently utilize them for other on demand instance request or spot instance request with respect to the risk of Service Level Agreement (SLA) promised to the customers. In this research, we address the problem of maximizing revenue for cloud provider. We try to solve the following research questions:

- With limited resources available, and considering the dynamic and stochastic nature of customers demand, how can the expected revenue be maximized through an optimal allocation of capacity to each pricing scheme.
- Moreover, with the unused resource from reservation contracts, how can we estimate the overbooking factor so that we utilize them to allocate for other instance requests to maximize revenue while respecting the SLAs promised to customers.

2. REFERENCES

http://aws.amazon.com/ec2/pricing/
Fault-Tolerant Workflow Scheduling Using Spot Instances on Clouds

[Extended Abstract]

Deepak Poola, Kotagiri Ramamohanarao and Rajkumar Buyya
Department of Computing and Information Systems,
The University of Melbourne,
Australia
deeapkc@student.unimelb.edu.au, {kotagiri,rbuyya}@unimelb.edu.au

ABSTRACT
A scheduling algorithm that schedules tasks on Cloud resources using two different pricing models (spot and on-demand instances) to reduce the cost of execution whilst meeting the workflow deadline is proposed. This algorithm is fault tolerant against the premature termination of spot instances (SIs) and also robust against performance variations of Cloud resources. Experimental results demonstrate that our heuristic reduces up to 70% execution cost as against using only on-demand instances.

Categories and Subject Descriptors
C.4 [Performance of Systems]: Fault tolerance; C.2.4 [Computer-Communication Networks]: Distributed Systems—Distributed applications

Keywords
Workflows, cloud, fault-tolerance, scheduling, spot instances

1. INTRODUCTION
A just-in-time and adaptive scheduling heuristic is presented here. The scheduling algorithm, for every ready task, evaluates the critical path and computes the slack time, which is the time difference between the deadline and the critical path time. The main motivation of the work is to exploit SIs to the extent possible within the slack time. As the slack time decreases due to failures or performance variations in the system, the algorithm adaptively switches to on-demand instances. The algorithm employs a bidding strategy and checkpointing to minimize cost and to comply with the deadline constraint. Checkpointing can tolerate instance failures and reduce execution cost, in spite of an inherent overhead.

2. PROPOSED APPROACH
The crux of the algorithm is to map tasks that arrive before the Latest Time to On-Demand (LTO) to SIs and those that arrive after the LTO to on-demand instances. LTO is the latest time the algorithm has to switch to on-demand instances to satisfy the deadline constraint.

Initially, CP and LTO are computed before the workflow execution. They are recomputed for all ready tasks during execution. When a new task is ready to be mapped, the algorithm tries to pick empty slots among the existing running instances. If there is no free slot it searches for a running instance that will be free before the task’s latest start time. If no such instance is found the algorithm instantiates a new instance considering its and performance characteristics.

2.1 Bidding Strategy
The proposed Intelligent Bidding Strategy takes into account the current spot price ($p_{spot}$), on-demand price ($p_{OD}$), failure probability ($FP$) of the previous bid price, $LTO$, the current time ($CT$), $\alpha$ and $\beta$. $\alpha$, as seen in Equation 1, dictates how much higher the bid value must be above the current spot price. $\beta$ determines how fast the bid value reaches the on-demand price. $FP$ of the previous bid is used as a feedback to the current bid price, the current bid price varies in accordance to the $FP$ adding intelligence to the bidding strategy. The bid price increases gradually with the workflow execution and as the $CT$ moves closer to the $LTO$. The bid starts around the initial spot price and ends closer to the on-demand price. Lower the value of $\alpha$, higher is the value of the bid w.r.t the spot price.

$$\gamma = (-\alpha(LTO - CT))/FP$$

$$bid = e^\gamma * p_{OD} + (1 - e^\gamma * (\beta * p_{OD} + (1 - \beta) * p_{spot}))$$ (1)

3. RESULTS AND CONCLUSION

Simulation results (Fig 1) show that cost reductions of upto 70% for our algorithm (CIB) are achieved under relaxed deadlines, when SIs are used. This work also demonstrates the use of checkpointing, which offers cost savings up to 14%. This works presents an adaptive scheduling heuristic and also an intelligent bidding strategy.
ABSTRACT
Internet is continuously growing in multiple dimensions and as it gets bigger, the number of devices using it as a means for communication has been magnified. Currently, Cloud Computing is referred to as a de-facto standard for dealing with Big Data that requires elastic computing and on-demand escalation and such potential can be leveraged for the sake of IoT applications, if it is defined and used properly. To feel the gap in this area, I have modified Aneka, which is a platform for Cloud Computing as a service, to address IoT needs and provide features such as load balancing, resource provisioning, and scheduling to IoT applications, besides serving other cloud-based applications.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Design

Keywords
Internet of things, cloud computing, aneka

1. INTRODUCTION
Internet of Things (IoT) is a new paradigm in computer science that aims to leverage numerous connectivity facilities to connect different “Things” together and form a ubiquitous network of interconnected nodes. The term “Thing” here refers to any device or object that can use Internet as a backbone to communicate with other network nodes at any time [1].

RFID tags, Wireless Sensor Networks (WSN), NFC-enabled devices, microchips with Wi-Fi connectivity, and many other hardware technologies can be used in the IoT context to create the desired network. Additionally, there should be a universal mechanism for identifying every node in the network, which is achieved by multiple software layers and protocols operating tightly with the hardware layer. IoT has recently gained a lot of attention from giant tech companies such as Apple, Google, Samsung, Cisco, IBM and also startups and research community as well, due to its potential and capabilities to be used in versatile applications, such as: Smart home, Healthcare, Supply chains/Logistics, Social applications, Disaster prediction and management, and many others [2].

2. ANEKA- A PLATFORM FOR IOT
One of the key requirements of IoT applications is the ability to use scalable computing power to process the data coming from different devices. By considering Cloud Computing features, using a cloud to address the above mentioned challenge seems plausible, but if there is no suitable integration mechanism between cloud providers and IoT applications, developers cannot use the complete power of cloud computing to meet their demands. As an example, Aneka is a Platform as a Service (PaaS) software package that handles resource provisioning, task allocation and scheduling, and many other cloud features such as load balancing, but it was not designed to handle IoT specific applications.

By using a standard IoT communication protocol developed by IBM, which is called MQTT, we extended the Aneka framework and added the support of IoT application processing to it, thus now, a developer can simply use the infrastructures provided by Aneka to solve his task using different cloud providers. This way, developers can put more time on their application’s business logic, without having to worry about underlying architecture and solutions [4].

3. OPEN CHALLENGES
Since IoT is a distributed environment and utilizes Internet as backbone for communication, a lot of challenges will arise in terms of security, privacy, and accountability. Furthermore, finding suitable approaches to uniquely identify and address any object in the IoT environment is a challenging issue, due to the increasing size of Internet-capable devices and different communication mechanisms used by different vendors. In addition, finding the best visualization and representation method to be used in IoT is still an open issue and needs further investigation [2,3].

4. REFERENCES
Energy Efficient Software Defined Clouds

[Extended Abstract]

Jungmin Son
The University of Melbourne
jungmins@student.unimelb.edu.au

ABSTRACT
In this paper, we discuss energy efficient data center for clouds through software defined clouds.

Categories and Subject Descriptors
C.2.4 [Computer-Communication Networks]: Distributed Systems

Keywords
Cloud computing, SDN, virtual network

1. INTRODUCTION
Cloud computing [1] has emerged rapidly in industry in the last few years. It substitutes conventional server rooms in service providers with its Infrastructure-as-a-Service (IaaS) model that enables service providers to use virtually created servers in cloud provider’s data center. Cloud computing also enables Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) models on top of IaaS which provides platforms or software for customers by provisioning underlying resources. In order to provide these services, cloud providers build multiple data centers holding thousands of physical machines and hundreds of switches connecting them.

Different customers have different requirements. Some customers would utilize clouds for compute intensive jobs, while others would use it for conventional web servers. As each of various requirements needs a different configuration, managing a data center is complex and difficult job.

Nonetheless, it is critical for cloud providers to manage virtual resources efficiently to ensure Software Level Agreements (SLAs) to their customers. Cloud providers must guarantee SLAs to make sure for service providers to provide sufficient Quality of Service (QoS) to their users. Various requirements by different service providers make it difficult, especially when their requirements dynamically change. As one of the key aspect in cloud computing is elasticity for fluctuating service usages, service providers are willing to scale their service dynamically, which results in various strategy for cloud providers such as migrating a virtual machine to another physical machine. On the other hand, it is important for cloud providers to consolidate virtual resources in order to manage physical machines more efficiently, so that they can make more profit as well as reduce energy consumption. In short, cloud data centers should be managed efficiently while ensuring SLAs for the customers.

One approach to manage virtual resources dynamically in cloud data centers while guaranteeing SLA compliance is Software Defined Clouds (SDC) [3]. SDC allows dynamic reconfiguration of virtual resources across physical machines to meet SLA requirements. It enables cloud resources to be dynamically configured and managed through software controller. SDC is consisting of multiple technologies: system virtualization, cloud computing, software-defined network, software-defined middleboxes and network virtualization. In this research, we look into various aspects of SDC and its components to make energy efficient cloud data center while ensuring SLAs.

One of the key elements in SDC is software-defined network (SDN). The main concept of SDN underlie to separation of control plane from data forwarding plane to enable controlling overall network by central software controller. By controlling network through software, it is easy to reconfigure the network topology which can make migration of virtual machines much easier. Also, it can be used to change the virtual network topology which can reduce latency or increase bandwidth required to ensure SLA. Network virtualization techniques are necessary to make it feasible.

2. EVALUATION METHOD
In order to evaluate energy consumption and performance of SDC, a simulation tool is designed based on CloudSim [2]. CloudSim is a cloud simulation framework widely used for research purposes. On top of CloudSim core architecture, we add additional components to support the physical network topology, virtual network environment, and SDN controller. This architecture includes support for different levels of network switches and middleboxes.

3. REFERENCES
Resource Provisioning and Scheduling Algorithm for Workflows on Clouds
[Extended Abstract]

Maria A. Rodriguez and Rajkumar Buyya
Cloud Computing and Distributed Systems (CLOUDS) Laboratory
mariaars@student.unimelb.edu.au, rbuyya@unimelb.edu.au

ABSTRACT
Cloud computing is the latest distributed computing paradigm and it offers tremendous opportunities to solve large-scale scientific problems. This paper proposes a resource provisioning and scheduling strategy for scientific workflows on Infrastructure as a Service (IaaS) Clouds. We present an algorithm based on Particle Swarm Optimization (PSO), which aims to minimize the overall workflow execution cost while meeting deadline constraints. Our heuristic is evaluated using various well-known scientific workflows of different sizes and the results show that it performs better than the current state-of-the-art algorithms.

Categories and Subject Descriptors
C.2.4 [Distributed Systems]: Distributed applications

Keywords
Cloud, cost-minimization, deadline, resource provisioning, scheduling, workflow

1. INTRODUCTION
Workflows are frequently used to model large-scale scientific problems that have ever-growing data and computing requirements and demand a high-performance computing environment in order to be executed in a reasonable amount of time. These workflows are modeled as a set of tasks interconnected via data or computing dependencies. The orchestration of these tasks onto distributed resources has been studied extensively over the years, focusing on environments like Grids and Clusters. However, with the emergence of new paradigms such as Cloud computing, novel approaches that address the particular challenges and opportunities of these technologies need to be developed.

We develop a cost-minimization, deadline-constrained heuristic for scheduling scientific workflows on Clouds. Our approach considers fundamental features of IaaS providers such as the dynamic provisioning and heterogeneity of unlimited computing resources as well as (Virtual Machine) VM performance variation. To achieve this, both resource provisioning and scheduling are merged and modeled as an optimization problem. PSO is then used to solve such problem and produce a schedule defining not only the task to resource mapping but also the number and type of VMs that need to be leased, the time when they need to be leased and the time when they need to be released. Our contribution is therefore, an algorithm with higher accuracy in terms of meeting deadlines at lower costs that considers heterogeneous resources that can be dynamically acquired and released and are charged on a pay-per-use basis.

2. PROPOSED SOLUTION
To define the PSO model, we need to establish the meaning and dimension of a particle. For the scheduling scenario presented here, a particle represents a workflow and its tasks; thus, the dimension of the particle is equal to the number of tasks in the workflow. The range in which the particle is allowed to move is determined in this case by the number of resources available to run the tasks. As a result, the value of a particle’s coordinate can range from 0 to the number of VMs in the initial resource pool. In this way, the particle’s position encodes a mapping of task to resources. Since the fitness function is used to determine how good a potential solution is, it needs to reflect the objectives of the scheduling problem. Based on this, the fitness function will be minimized and its value will be the total execution cost associated to the schedule derived from the particle’s position. We omit the algorithm that generates a valid schedule form a particle’s position due to space constraints.

Because of the elasticity and dynamicity of the resource acquisition model offered by IaaS providers, there is no initial set of available resources we can use as an input to the algorithm. Instead, we have the illusion of an unlimited pool of heterogeneous VMs that can be acquired and released at any point in time. Consequently, a strategy to define an initial pool of resources that the algorithm can use to explore different solutions and achieve the scheduling objective needs to be put in place. Such strategy needs to reflect the heterogeneity of the VMs and give PSO enough options so that a suitable particle (i.e. solution) is produced. To keep the search space from becoming too large, we propose the following scheme. Let \( P \) be the set containing the maximum number of tasks that can run in parallel for a given workflow; then the initial resource pool that PSO will use to find a near-optimal schedule will be comprised of one VM of each type for each task in \( P \). Our algorithm will then select the appropriate number and type of VMs to lease from this resource pool. In this way, we reflect the heterogeneity of the computing resources and allow the algorithm to execute all the tasks that can run in parallel to do so.

3. EXPERIMENTS AND RESULTS
The simulation experiments conducted with four well-known workflows show that our solution has an overall better performance than the state-of-the-art algorithms, SCS [1] and ICPCP [2]. In every case in which ICPCP fails to meet the application’s deadline, our approach succeeds. Furthermore, our heuristic is as successful in meeting deadlines as SCS, which is a dynamic algorithm. Also, in the best scenarios, when our heuristic, SCS and ICPCP meet the deadlines, we are able to produce schedules with lower execution costs.

4. REFERENCES
Virtual Machine Customization and Task Mapping Model for Efficient Allocation of Cloud Data Center Resources

[Extended Abstract]

S. F. Piraghaj, R. N. Calheiros, R. Buyya
Cloud Computing and Distributed Systems (CLOUDS) Laboratory
Department of Computing and Information Systems
The University of Melbourne, Australia
s.fotuhipiraghaj@student.unimelb.edu.au, {rnc, rbuyya}@unimelb.edu.au

ABSTRACT

Although there has been significant research efforts in decreasing the energy consumed by data centers, most approaches lack the analysis of real cloud backend traces and their variance in the proposed solutions. To counter this issue, this paper proposes a new energy-aware methodology for mapping groups of tasks to customized virtual machine (VM) types based on the predicted task resource usage obtained from the analysis of the historical data. The objectives of this approach are efficient resource allocation and energy consumption improvement via the provisioning of a smaller number of efficiently utilized virtual machines.

Categories and Subject Descriptors
I.5.3 [Clustering]: Algorithms; Similarity measures; H.2.8 [Database Applications]: Data mining; D.4.8 [Performance]: Modeling and prediction

Keywords
Cloud Computing, Energy Efficiency, Workload Characterisation, Virtualization, Cluster Analysis, Data mining

1. INTRODUCTION

With the fast growth in high computational power demand, there is a growing need for large scale data centers and this leads to a shift towards cloud computing. However, energy usage of the large scale data centers has always been a concern for cloud providers as an average data center energy consumption is equal to thousands of households. In this respect, there has been an increased effort in decreasing the energy consumed in the data centers.

Most of the previous studies in energy efficiency area lack the analysis of real cloud backend traces and the effect of variance of the cloud workload on the proposed solutions. To counter this issue, this paper proposes a new architecture for cloud resource allocation that maps groups of tasks to customized virtual machine (VM) types based on the predicted task resource usage obtained from the analysis of the historical data.

The objectives of the proposed architecture are efficient resource allocation and energy consumption improvement via the provisioning of a smaller number of efficiently utilized virtual machines. In order to decrease resource wastage, VM configurations are defined by considering the workload characteristics in terms of average CPU, memory and disk usage for each group of tasks. Experimental results showed a considerable improvement in the average CPU utilization of active hosts compared to our baseline approach where allocation is based solely on amount of requested resources. In addition, the total data center power consumption decreases considerably in comparison to the same baseline.

The data set used in this paper is derived from the second version of the Google cloud trace log collected during a period of 29 days. The considerable gap between the actual reported resource usage and the requested amount for task execution shows that resource allocation in data centers experiencing load patterns similar to the one described by the Google traces can be improved.

The efficient allocation of resources is crucial in data centers since resource wastage results in a higher energy consumption. In order to tackle this problem, the focus of the proposed architecture is on efficient allocation of physical resources. In this respect, we leverage virtualization technology, which is the key feature introduced in cloud data centers. This technology enables efficient utilization of resources and load balancing via migration and consolidation of workloads in cloud environment. In addition, a considerable amount of energy can be saved with virtual machine migrations from underloaded servers by putting them in a lower power state. The contributions of this paper are as follows:

- It presents an architecture for efficient allocation of requests on data centers that reduces the infrastructure's energy consumption;
- It presents an approach, applied to the proposed architecture, to group tasks.
- We present experiments showing that our proposed approach results in less number of servers, which in turn results in less energy consumption in the data center.
Integrated Framework for Cloud-based Interoperability between Social Network Websites

[Extended Abstract]
Safiollah Heidari, Rajkumar Buyya
Department of Computing and Information Systems, The University of Melbourne
sheidari@student.unimelb.edu.au
rbuyya@unimelb.edu.au

ABSTRACT
In this study, we try to find a framework for interoperability problem in the next generation of cloud-based social networks that are not only under control of a particular provider, but also more than one provider.

Categories and Subject Descriptors
C.2 [Computer-Communication Networks]: Distributed Systems; D.2.12 [Interoperability]: Distributed applications – Distributed databases

Keywords
Cloud computing, social networks, interoperability

1. INTRODUCTION
Cloud Computing and Social Network Sites (SNS) are among the most controversially discussed developments in recent years. The opportunities of using powerful computing resources on demand via the web are considered as a possible driver for the growth around the world. However, there are also critics arguing that economic, social and technical risks prevail or even dismiss the potentials of Cloud Computing and SNS. In this study, we try to find a solution for interoperability problem in the next generation of cloud-based social network websites.

2. CHALLENGES
In recent years, social network websites have attracted a lot of attention and grow rapidly. Each of these websites has been established with a specific purpose and offers a range of particular services. For example, some of them are for communication and entertainment, some for scientific and educational services, some for health and tourism, etc. Hence, people, with different needs, register to these websites to use their services. But there are some challenges as follow that should be considered.

- Each SNS has its own policies and procedures and users have to go through certain processes which are specific to each of them.
- Users cannot use different SNSs simultaneously, under a unified system.
- If a user needs to transfer his/her information from one SNS to the other one, with a different provider, he/she cannot do so.
- By growing various services on SNSs, if a user wants to integrate different services on a specific platform, it is not easily possible again.
- Possibility of exchanging data and information between different SNSs, that are based-on clouds, will become inevitable in near future.

3. SOLUTION
There are some solutions to overcome the challenges mentioned above:

- A standard framework provides a unified platform for integrating various SNSs in software and platform level with respect to flexibility and maintainability issues.
- Data sync operation is needed at both sides of each relationship that can be done by a data mediator with automatic or on-demand methods.

4. REFERENCES
ABSTRACT

Big data refer to various and voluminous data sets which are continuously generated at high velocity that are difficult to process using traditional data processing tools. Big data applications are developed to analyze and find potential value and “secret” from big data to enable government, enterprises and academic make wise decisions, i.e., product trends prediction, disaster prediction and management, business strategy making. Big Data applications require large amount of Cloud resources, i.e., CPU, RAM, Storage, etc., to process big data. How to efficiently schedule resources from multiple Clouds that can on the one hand satisfy specified Quality of Service (QoS) requirements of applications and on the other hand reduce resource consumption and cost is an urgent research problem raised for Big Data computing. To solve this, we design a Service Level Agreement (SLA) Based Resource Scheduling Platform for Big data applications.

Categories and Subject Descriptors
C.2.4 [Distributed Systems]: Distributed applications; D.2.2 [Software Engineering]: Design Tools and Techniques; F.1.1 [Theory of Computation]: General

Keywords
Big Data, cloud computing, sla, resource scheduling

1. INTRODUCTION

1.1 Cloud Computing

Cloud Computing becomes the fifth utility to provide resources and services to satisfy daily living needs after other four utilities: water, electricity, gas and telephony [1]. Cloud computing conveniently provides heterogeneous resources to users from various domains in pay as you go model which allows users focus on their core business without heavily invest on underlying infrastructure and platform.

1.2 Big Data

Big data has been one of the current and future research frontiers that has drawn huge attention from researchers. Nowadays, huge amount of data are generated every day from different sources, i.e., media, sensors, satellites, social websites, and mobile devices, which requires large amount of storage resources and computing resources to store and process them. Big data has the characteristic of 3V, which are: variety, velocity and volume [2]. Analyzing and gaining insight of big data will revolutionize many fields, i.e., industry, scientific research, government administration [3].

Big data applications consumes huge amount of Cloud resources, CPU, RAM, and storage during data analytic process. For Big Data applications scheduling, we need to provision resources that can not only fulfill functional requirements of applications but also can satisfy QoS requirements of applications, i.e., cost, deadline, security, etc. Since resource intensive Big Data applications require large amount of resources during their execution, the scheduling mechanism should be scalable to lease and release large amount of resources elastically due to real-time demands. How to efficiently and dynamically schedule and provision Cloud resources for Big Data applications is the research problem that we want to solve by designing a SLA based Resource Scheduling platform.

1.3 Platform Design

SLA based Resource scheduling platform for Big Data applications composes of five components: User Interface, Performance Estimation and Prediction component, Resource Scheduling component, Resource Allocation component and Resource Controlling component. The user interface component accepts applications specification from users. Performance Estimation and Prediction component matches application specification with cloud providers’ profiles to estimate whether user requests can be accepted or not. If so, SLA is built to constraint resource provision process. Resource Scheduling component selects the best suitable resources for applications that can not only satisfy QoS requirements but also minimize cost. Resource Allocation component allocates resources for applications’ execution. Resource Controlling component is responsible for dynamically control resources based on application running demands.

2. REFERENCES


Virtual Machine Allocation Policies against Co-resident Attacks in Cloud Computing
[Extended Abstract]

Yi Han
Department of Computing and Information Systems
The University of Melbourne
andrew.hanyi@gmail.com

Jeffrey Chan
Department of Computing and Information Systems
The University of Melbourne
jeffrey.chan@unimelb.edu.au

Tansu Alpcan
Department of Electrical and Electronic Engineering
The University of Melbourne
tansu.alpcan@unimelb.edu.au

Christopher Leckie
Department of Computing and Information Systems
The University of Melbourne
caleckie@unimelb.edu.au

ABSTRACT
While the services-based model of cloud computing makes more and more IT resources available to a wider range of customers, the massive amount of data in cloud platforms is also becoming a target for malicious users. Previous studies show that attackers can co-locate their virtual machines (VMs) with target VMs on the same server, and obtain sensitive information from the victims using side channels. This paper investigates VM allocation policies and practical countermeasures against this kind of attack by developing a set of security metrics and a quantitative model. A security analysis of three commonly used VM allocation policies reveals that the server’s configuration, oversubscription and background traffic have a large impact on the ability to prevent an attacker from co-locating with the targets. If the servers are properly configured, and oversubscription is enabled, the best policy is to allocate new VMs to the server with the most VMs. Based on these results, a new strategy is introduced that effectively decreases the probability of achieving co-residence. The proposed solution only requires minor changes to current allocation policies, and can be easily integrated into existing cloud platforms to mitigate the threat of co-resident attacks.

Categories and Subject Descriptors
K.6 [Management of Computing and Information Systems]: Security and Protection; D.2.8 [Software Engineering]: Metrics – performance measures

Keywords
Cloud computing security, co-resident attack, virtual machine allocation policy

1. INTRODUCTION
The emergence of cloud computing has fundamentally changed the deployment and usage of information technologies. More and more IT resources, like software applications, operating systems, and even network infrastructure, are now delivered as services and made accessible to a wide range of customers. However, while regular users benefit from the advantages brought by cloud computing, malicious users are also targeting the growing amount of data in cloud platforms, which creates a major potential security risk.

This paper focuses on a novel type of threat: the co-resident attack [1]. In cloud computing environments, in order to maximise the utilisation rate of hardware platforms, it is common practice that the virtual machines (VMs) of different users run on the same physical server (i.e., these VMs are co-resident), and are logically isolated from each other. However, malicious users can circumvent the logical isolation, and obtain sensitive information from co-resident VMs. If cloud providers cannot ensure data confidentiality and lose the trust from users, the future of cloud computing will be jeopardised. Therefore, it is crucial to find effective and practical countermeasures against this kind of threat.

One way to encounter this kind of threat is to eliminate the side channels, which is the approach taken by most previous work. However, the proposed methods require substantial changes to existing platforms, and hence are impractical and not suitable for immediate deployment.

In this paper, we approach the problem from a different perspective. Before the attacker is able to extract any useful information, they first need to co-locate with their targets. If we can find a practical way to decrease the possibility of achieving co-location, then the threat of this attack can be mitigated. Specifically, we focus on the VM allocation policy, since this is one important factor the cloud provider can control that will influence the possibility of co-location. Our contributions include: (1) we define three security metrics for assessing attacks, and quantitatively model these metrics under different allocation policies; (2) we conduct extensive experiments to verify the models; (3) we find that if the servers are properly configured, and oversubscription is enabled, the best policy is to allocate new VMs to the server with the most VMs; and (4) we propose a new VM allocation strategy that considerably increases the difficulty of achieving co-residence.

2. REFERENCES
The influence of personality traits on online information behaviour during political campaigns

Miguel Wood
University of Melbourne
mwood3960@gmail.com

Simon Milton
University of Melbourne
simon.milton@unimelb.edu.au

Graeme Shanks
University of Melbourne
gshanks@unimelb.edu.au

ABSTRACT

The adoption of social analytics heralds the arrival of a new paradigm in political campaign practice, driven by both behavioural science and social networks. Yet these domains have historically been in tension with opposite perspectives on political behaviour. The increasing focus on online social networks offers an opportunity to synthesise complementary research traditions on political behaviour to better understand the relationship between individual personality and information behaviour. We propose a model based on core dispositional personality traits, political participation as information consumption and information diffusion theory to explain online information behaviour during election events. Information diffusion requires users to perform several actions in combination - engage, rate content, comment, share and interpret information received. The research model can be used in empirical research to explore how personality type guides interaction with and diffusion of political information across online social networks. The learned behaviors' and direct responses of participants may help political campaigns in the design of future behavioural targeting strategies that activate social network effects to achieve campaign goals.

Categories and Subject Descriptors

H.1.1 [Systems & Information Theory]: Information theory; H.1.2 [User/Machine Systems]: Human factors, software psychology, social networks; J.4 [Social & Behavioural Science]: Sociology

Keywords
Personality traits, information behaviour, online social networks, information diffusion, political campaigns

1. INTRODUCTION

The adoption and investment in sophisticated infrastructure platforms by political campaigns has enabled faster decision-making, improved segmentation and targeting of voters and improved cost effectiveness to help secure election victory. A cornerstone capability has been the use of knowledge discovery and data-mining (KDD) tools for generating actionable intelligence to support voter categorisation, mobilisation, and persuasion [1, 2]. Pervasive changes in interconnectivity translated into a step change for digital campaigning during the recent 2012 US election cycle. Increased maturity and sophistication in development operations, data integration, predictive analytics and online social networks has yielded greater precision in voter targeting and field operation effectiveness [3]. A significant shift in data-driven insights is detectable with the move towards dynamic, real-time monitoring based on unstructured data streams accessible via online social networks.

So what does the future hold for political campaigns? The relentless pace of change and need to remain competitive sees a growing focus at the intersection of behavioural science and online social networks. Priority areas include enhanced social targeting, real-time analytics to overtake polling, and genuine digital integration [3]. With the availability of information technology enabled social networks we revive a once regularly studied area of political science, the role of personality traits in predicting political interest, knowledge, consumption and participation. As our society embarks on a journey through the Digital Age a window of opportunity is emerging to redesign channels for mass communication and democratic participation.

2. RESEARCH MODEL

The research model uses information diffusion theory to demonstrate the relationship and significance of the influence of personality traits on online information behavior during political campaigns. The use of pervasive online social networks enables observation of the spread of information at scale and associated network effects. The research model draws together constructs of personality traits and online interaction (see Figure 1 below).

3. REFERENCES


A conceptual model for analysing informal learning in online social networks for health professionals

[Extended Abstract]

Xin Li
Department of Computing and Information Systems
xinl5@student.unimelb.edu.au

Kathleen Gray
Department of Computing and Information Systems
kathleen.Gray@unimelb.edu.au

ABSTRACT
Online social networking (OSN) provides a new way for health professionals to communicate, collaborate and share ideas with each other for informal learning on a massive scale. It has important implications for ongoing efforts to support Continuing Professional Development (CPD) in the health professions. However, the challenge of analysing the data generated in OSNs makes it difficult to understand whether and how they are useful for CPD. This paper presents a conceptual model for using mixed methods to study data from OSNs to examine the efficacy of OSN in supporting informal learning of health professionals. It is expected that using this model with the dataset generated in OSNs for informal learning will produce new and important insights into how well this innovation in CPD is serving professionals and the healthcare system.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics − complexity measures, performance measures

Keywords
Social network analysis, social learning, networked learning, informal learning, health professional education.

1. INTRODUCTION
Staying up-to-date and delivering best evidence-based care is crucial for health professionals. They need to be lifelong learners as medical knowledge expands and changes rapidly. Some evidence suggests that current CPD approaches (e.g., educational courses, lectures) are not efficient or effectively meeting this need and thus are inadequate to improve the safety and quality of patient care [1].

This and wider Internet use factors have led to the emergence of a more informal approach to CPD making use of OSN [2]. OSN plays an important role in supporting the CPD of health professionals as it allows health professionals to stay up-to-date and knowledgeable of the current literature; it has been increasingly used for learning from experts, sharing medical knowledge, discussing practice management challenges and clinical issues. For example, one OSN site in the US, sermo.com has over 200,000 members from nearly 70 medical specialties.

However few efforts have been made to analyse such OSN data for insights about the CPD they are supporting; the scarce evidence about this phenomenon makes it difficult to identify and understand how to increase CPD efficiency and effectiveness for individual health professionals through OSN, and thereby improve patients’ outcomes [1]. One factor impeding such research is the difficulty of making sense of their large and complex datasets. This paper proposes a conceptual model that can be applied to analysing these data, to understand more about the informal learning occurring on OSNs for health professionals.

2. CONCEPTUAL MODEL
As shown in Figure 1, we propose that the data about informal learning of health professionals on OSNs can be analysed using SNA, CA and survey method, along three dimensions: learning interaction, learning process and learning outcome.

3. CONCLUSION
A conceptual model for analysing informal learning occurring in OSN for health professionals can enable learners, and those who design and manage learning, to make better use of the data they generate. Our model describes the essential components of learning interaction, process and outcome based on relevant health literature and theories. For future studies, we will evaluate this model by applying it to the analysis of a large online medical community used by more than 10000 health professionals during the period 2007-2012. Ongoing research, using our model with the dataset from any active OSN for health professionals’ informal learning, is expected to refine the model and to produce new and important insights into how well this innovation in CPD is serving professionals and the healthcare system.

4. REFERENCES
[1] Institute of Medicine, Redesigning Continuing Education in the Health Professions. 2010, National Academies Press: Washington, DC.
Task Complexity and Codification

[Extended Abstract]

Marion Zalk
University of Melbourne
mzalik@student.unimelb.edu.au

Rachelle Bosua
University of Melbourne
rachelle.bosua@unimelb.edu.au

Rajeev Sharma
University of Wollongong
rajeev@uow.edu.au

ABSTRACT
The importance of knowledge and knowledge creation within an organisation is widely recognised, however the process of creating and managing knowledge is not well understood [7]. This research will address a key element in the knowledge creation process, specifically knowledge sharing behaviour within organisations. This act of knowledge sharing is a key precursor for the actual usage of knowledge [2]. Knowledge transfer has two components: a source and a recipient [6]. This research is concerned with the source, specifically with his/her perceived task complexity of codification. The issue here is twofold as we have two task complexities - the task complexity of the primary work task and the secondary task of codifying the primary task. A quantitative research method is employed in order to consider the relationship between the complexities of the different tasks.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics – complexity measures, performance measures

Keywords
Task complexity, codification, externalization, experiment, knowledge management

1. INTRODUCTION
The concept of knowledge is broad and abstract. Many authors define knowledge by creating a hierarchy of data, information and knowledge. Data refers to raw numbers and facts and is combined to create information (processed data). Information is then combined to create knowledge [1]. Knowledge can refer to a competence, a skill, an experience, a value judgment and/or social networks. A guiding definition of knowledge is ‘a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the mind of knowers [4]. However as knowledge is often personalized it is important that it is expressed in a manner that is interpretable to others. Stores of information provide little value, only information actively processed in the individual through a process of reflection, enlightenment or learning is valuable [1]. Knowledge can be described along the dimensions of tacit and explicit [5]. Tacit knowledge is tied to the senses, embodied skills, individual perception, physical experiences, rules of thumb and intuition [8]. It is therefore personal and is difficult to articulate, complex to understand and may even be subconscious [5]. Knowledge is explicit when it is communicated or codified (put on paper, formulated into sentences or captured in a drawing/s) [8]. This knowledge can either be broad or specific. Broad knowledge is often publically available and independent of particular events whereas specific knowledge is context specific so that it meaningful in a situation and with a specific focal point [9].

Nonaka and Takeuchi [7] describe how tacit and explicit knowledge are converted. Of interest to this research is externalization. Externalization is the articulation of best practices or lessons learned. With knowledge transfer there are 2 tasks that need to be discussed. Initially there is the primary work task and then the task of documenting the knowledge required to complete the work task. To easily differentiate these will be referred to as work task (WT) and documenting task (DT).

Task complexity has been discussed in a least three distinct bodies of literature, namely information processing and decision-making literature; task and job design literature and goal setting research literature. On one extreme task complexity can be considered purely as a psychological experience and on the other extreme purely from the task characteristics. In the middle is consideration of both the task and the person characteristics [3].

2. REFERENCES
Online Medical Consultation: A review of current practices

[Extended Abstract]

Ibrahim AL-Mahdi
University of Melbourne
ibrahim.al@unimelb.edu

ABSTRACT
Online Medical Consultation (OMC) has been offered by many providers since the beginning of this century. There is a growing trend towards using the internet as a source of health information as well as for medical consultation. This study performs an empirical analysis on a group of existing OMC sites to explore their major themes, modalities, geographical coverage, and claimed value. These features have been illustrated for a better understanding of the promise on which these services operate. The study found out that OMC is a growing phenomenon featuring several interaction modalities, serving various medical consultation purposes, and accessible to consumers throughout the world. The contribution of this work is to synthesise a list of features and benefits of OMC services and to understand this phenomenon within the wider context of the rise of Internet services.

Categories and Subject Descriptors
K.4.1 [Public Policy Issues]: Computer-related health issues; J.3 [Life and Medical Sciences]: Health

Keywords
Consultation, telemedicine, remote consultation

1. INTRODUCTION
People seek information, read news, do banking, trade, play, socialize, and learn online. Seeking health advice and treatment online is no exception to that, where many have now added internet to their "personal health toolbox" to help them figure out their medical issues and seek treatment [1]. A simple internet search of ‘online doctor’, or ‘online medical consultation’, returns hundreds of links for sites ranging from free ask-the-doctor sites to highly prestigious sites with diagnostic tools and multi-interactive options. Academic reviews of telemedicine have also cited numerous advantages for patients such as increased accessibility to health services, reduced travel and waiting time to see a doctor, and reduced cost [2-4]. The aim of this paper is to explore the current range of OMC services, to analyse their main features, and to summarize the benefits and successes of OMC as collectively presented by their providers.

2. METHOD
An empirical analysis to examine the themes of multiple OMC web sites was conducted on March 2014. The selected sites appeared among Google’s first hundred results when searching for the term (‘online consultation’ AND (medical or Health)). Exclusion was made to web sites which are not providing a private and online channel for conducting a patient-doctor consultation. Most of the exclusions occurred for web sites whose main purpose is online pharmacies, public and generic health information exchange, advertising, wellbeing only advice, online booking, or telephone-only based consultations.

Relevant web sites were inspected to explore the supported modalities of the consultations, the intended purpose of the consultation, the payment options, the medical disciplines, and the geographical coverage. Some data were sourced directly from these web sites and some were requested by email or by analyzing associated media reports.

3. FINDINGS AND CONCLUSION
According to the sample sites, the study revealed that OMC services proliferate with a rate of 150% each five years since year 2000. The modalities analysis showed that the main medium for communication was the private portal (68%) which is usually used as an entry point to provide basic patient details followed by consultation using other modality. Video conference is the second most popular modality (61%) due to its known value of communicating both voice and picture simultaneously. Telephone as well as email and private forums were also utilized. Some of the services are supported with special smart phone apps.

OMC costs vary from very low rates to several hundreds of US dollars per consultation service. The covered OMC sites are located at six countries where the US comes on top of the list. It was found that the examined OMC services were mainly for the purpose of diagnosis, wellbeing advice, nonprescription treatment and a fewer number were providing prescription based treatment. Most of the investigated sites are not restricted to any medical discipline but some stated that certain cases may be rejected due the need for physical examination.

In terms of benefits and value, OMC web sites promise to enable better access (at any time and from any place), more choices, less cost, less time to be consumed for conducting OMC consultations.

Findings’ details will be provided at the presentation.

4. ACKNOWLEDGMENTS
My special thanks to my supervisors Dr. Kathleen Gray, and Dr. Reeva Lederman for their support. A special recognition goes towards my scholarship sponsor, the University of Melbourne, and to IBES.

5. REFERENCES
Performance evaluation of spectral unmixing method based on different ways of training data selection

[Extended Abstract]

Liyan Liu
Department of Computing and Information System
liyanl@student.unimelb.edu.au

Andrey Kan
Walter and Eliza Hall Institute of Medical Research
akan@csse.unimelb.edu.au

Christopher Leckie
Department of Computing and Information System
caleckie@unimelb.edu.au

ABSTRACT
Fluorescence multi-channel imaging suffers from inherent problem of spectral bleed-through. Spectral unmixing method has been widely adopted to solve that problem. However, almost all the methods of evaluating the performance of spectral unmixing are based on the reference objects which are unfortunately not available in most applications. Moreover, there is little discussion about how to select training data as an initialization of spectral unmixing. So our purpose is to address the challenge of missing reference objects by proposing a new performance evaluation approach.

Categories and Subject Descriptors
I.4 [Image processing and computer vision]: General; J.3 [Computer Application]: Life and medical sciences—Medical information systems

Keywords
Fluorescence time-lapse imaging, spectral unmixing, training data selection, performance evaluation

1. INTRODUCTION
Fluorescence time-lapse imaging is becoming an increasingly important tool in cell biology. However, the process of image acquisition microscopy setup leads to distortion of the original signal. Prominent acquisition effects include uneven illumination and spectral bleed-through, also called cross-talk. Here we mainly consider cross-talk because we assume that the illumination has been corrected. In order to computationally reduce that effect, spectral unmixing has been commonly adopted since it can separate each original signal from the recorded mixtures [1].

Based on the linear mixing model of spectral unmixing, we aim to recover mixing matrix and unmixed images simultaneously. However, the estimation of mixing matrix requires training data in advance. Unfortunately, there is no clear argument with regards to how to obtain this training data from a set of images. In some cases, it is obtained based on the reference object: a part of image that is known to be fluorescent in one channel but non-fluorescent in another channel, which is not feasible in most applications [2]. Moreover, it brings about another issue: how to evaluate the performance of spectral unmixing without reference objects. Hence, we aim to solve those two related issues.

2. METHODS
2.1 Spectral unmixing model
The process of image acquisition is illustrated in figure 1 below. The acquisition effects refer to spectral bleed-through, uneven illumination and camera noise. After removing camera noise and correcting uneven illumination, we establish such a relationship between real images and mixed images as [1]

\[ I_{\text{real}} = A \ast I_{\text{mix}} + R \] (1)

where A refers to the mixing matrix, and R indicates the additive white Gaussian noise.

Figure 1: Acquisition effects in fluorescence time-lapse imaging system

2.2 Performance evaluation
First we should take training data selection into account because it is a basis of initialization of mixing matrix. The selection can be focused on random pixels or some pixels with the same feature such as intensity. In the absence of reference objects, it becomes difficult to tell whether the performance of spectral unmixing is good or not. So we design an indirect ground truth dataset based on the biology of FUCCI reporter. The underlying truth is that the intensities during some intervals should be flat after unmixing.

3. CONCLUSIONS
In summary, this paper has discussed two related issues: training data selection and performance evaluation of spectral unmixing in the situation of no reference objects. Then we have proposed new methods to tackle such problems.

4. REFERENCES
Retinal Cross Sectional Layer Segmentation using Optical Coherence Tomography

[Extended Abstract]

Md Akter Hussain
Department of Computing and Information Systems,
The University of Melbourne
hussain@student.unimelb.edu.au

Alauddin Bhuiyan
Commonwealth Scientific and Industrial Research Organization (CSIRO) Perth WA 6014, Australia
alauddin.bhuiyan@csiro.au

Kotagiri Ramamohanarao
Department of Computing and Information Systems,
The University of Melbourne
kotagiri@unimelb.edu.au

ABSTRACT
In this paper, we demonstrate an automated method for the layer segmentation of the retinal OCT image for reliable and repeatable analysis for predicting diseases like glaucoma, AMD.

Categories and Subject Descriptors
I.4 [Image Processing and Computer Vision]: Morphological; J.3 [Life and Medical Sciences]: Medical information systems

Keywords
OCT, segmentation, retina, pattern recognition

1. INTRODUCTION
Optical Coherence Tomography (OCT) is a well-established and noninvasive method which acquires the internal structure of biological tissue like retina, blood flow etc. Spectral Domain OCT is most usable image among available because of its acquisition speed and sensitivity of imaging quality. OCT of retina is the direct cross-sectional image of retinal anatomy which allows the early diagnosis and more sensitive monitoring of a variety of retinal and optic nerve head diseases, such as glaucoma, macular degeneration, and macular edema. Figure 1 shows an OCT image of retina with its 8 layers. OCT technology is an active research since 1991 and retinal OCT image segmentation has only been more fully explored since 1995. Segmentation remains one of the most difficult and at the same time most commonly required steps in OCT image analysis. No typical segmentation method exists that can be expected to work equally well. Segmentation is a challenging task due to intrinsic speckle noise, intensity pattern, low optical contrast and motion artifacts.

2. PROPOSED METHOD
We use preprocessing for smoothing image, directional intensity variation and relative position of layers to find out layer of retinal. Our system can detect RNFL, ONL, OS, and RPE layer. First, we smooth OCT image using median filter and Gaussian filter as a preprocessing step. After that, A-scan based search technique is employed for finding RNFL top boundary where intensity variation from black to white and intensity value are used as parameters for canny edge point. This will give broken pixels which will connected by canny edge and slope of line. RNFL layer position and similar approach are used for finding RNFL bottom, OS Top and RPE Bottom Boundary. Then OTSU clustering (n=2) is used to find RPE top layer between OS Top and RPE Bottom Boundary layers. Then using OS top and RPE bottom position and using above approach, we find out ONL layer. Figure 2 shows our system generated output with segmenting mentioned layer. We are currently exploring methods for determining accurate and connected line for every layer.

3. CONCLUSION
The goal of my research is to develop methods for OCT image analysis that can help determine various disease correlations with retinal layer intensity reflectivity or width of layer. Such as reducing ISe band with age and due to early AMD.

4. REFERENCES
An Effective Automated System for Grading Severity of Retinal Arteriovenous Nicking in Colour Retinal Images

[Extended Abstract]

Pallab Kanti Roy, Uyen T. V. Nguyen, Alauddin Bhuiyan, Kotagiri Ramamohanarao
royp@student.unimelb.edu.au, thivun@csse.unimelb.edu.au,
alauddin.bhuiyan@csiro.au, kotagiri@unimelb.edu.au

ABSTRACT
In this paper, an effective method is proposed for the analysis of retinal venular widths to automatically classify the severity level of Retinal Arteriovenous nicking.

Categories and Subject Descriptors
I.4 [Image Processing and Computer Vision]: Morphological; J.3 [Life and Medical Sciences]: Medical information systems

Keywords
Retinal image, arteriovenous nicking, pattern recognition

1. INTRODUCTION
Retinal Arteriovenous (AV) nicking is a precursor for hypertension, stroke and other cardiovascular diseases [1]. In colour retinal images, AV nicking appears as a decrease in venular caliber at both side of an artery vein cross-over point (Fig. 1).

Figure 1: Examples of AV crossing (a) AV nicking crossing (b) normal AV crossing

2. METHODOLOGY
First we automatically segment retinal vessels and detect AV cross-over points (Fig. 2(a)). Then based on the morphological features and colour information of the retinal vessel we detect vein (Fig. 2(b)). After that, we refine vein edge using gradient profiling and compute it’s width. In case of AV nicking, there is a decrease in the venular widths near the crossover point compared to the mean width of the full vein segment. Based on this property, we create our feature set to classify AV nicking. Let, $W_i$ represent the vein widths which are sorted based on the Euclidean distance from the crossover point and $C_{rn}$ (Fig. 2(c)) presents the mean width of first $n$ centre points starting from the crossover point as

$$C_{rn} = \frac{1}{n} \sum_{i=1}^{n} W_i.$$

If $c$ (Fig. 2(c)) is the middle point of the vessel centreline then width distribution of the overall vessel is computed as follows:

$$MW = \frac{1}{s} \sum_{i=c-s/2}^{c+s/2} W_i, s \text{ is empirically set to 20.}$$

The difference between $C_{rn}$ and $MW$ at different scale is used to classify AV nicking.

$$F_n = \frac{MW - C_{rn}}{MW}, \text{where, } n = [10, 20, 30]$$

Here, the difference between $MW$ and $C_{rn}$ is normalized by $MW$, which makes our features less sensitive to the variation of vein widths. We train a Random Forest (RF) classifier using these features to classify the severity of AV nicking.

Figure 2: (a) Original Image (b) detection of vein (c) width of the vein segment. Here $C_{r10}$, $C_{r20}$ and $C_{r30}$ represent mean width of the first 10, 20 and 30 center points from the crossover point. $MW$ represent the mean width computed in 20 pixels window from the middle (represented by C) of vein centerline.

3. RESULT
We have analyzed 47 color retinal images. From these images, 93 detected crossover points are selected to evaluate the performance of the proposed method. We compare the detection accuracy of our method with a recently published AV nicking classification method. Our proposed method shows 64.51% classification accuracy in-contrast to the reported classification accuracy of 49.46% by the current method [1].

4. REFERENCES
Shooting the Breeze Involves neither Shooting nor a Breeze: Predicting the Compositionality of Multiword Expressions

[Extended Abstract]

Bahar Salehi
Dept. of Computing and Information Systems, The University of Melbourne
NICTA Victoria Research Laboratory
Supervisors: Timothy Baldwin, Paul Cook
bsalehi@student.unimelb.edu.au

ABSTRACT
In this paper, we propose a simple, language independent and highly effective method for measuring the compositionality of multiword expressions (MWEs). The proposed method uses the translations of the MWE and its components.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]

Keywords
Multiword expressions, string/distributional similarity

1. INTRODUCTION
A multiword expression (MWE) is any combination of words with lexical, syntactic or semantic idiosyncrasy, in that the properties of the MWE are not predictable from the component words. In this paper, we focus on the semantic idiosyncrasy, also known as non-compositionality. As an example, shoot the breeze has semantic idiosyncrasy, since the meaning of “to chat” in usages such as It was good to shoot the breeze with you cannot be predicted from the meanings of the component words shoot and breeze.

To predict the degree of compositionality of an MWE, we require a way to measure the semantic similarity of the MWE with its components. Our hypothesis is that compositional MWEs are more likely to be word-for-word translations in a given language than non-compositional MWEs. Hence, if we can locate the translations of the components in the translation of the MWE, we can deduce that it is compositional (Three examples are shown in Table 1). We investigate two different similarity measures to compare the translations of components with the translation of the MWE: string similarity [1] and distributional similarity [2]. Our second hypothesis is that the more languages we use as the basis for determining translation similarity between the MWE and its component words, the more accurately we will be able to estimate compositionality.

Figure 1 provides a schematic outline of our method.

Among the string similarity measures that we experimented with, longest common substring and Smith Waterman algorithm were found to be superior to edit distance-based (Levenshtein) methods. Our results show that using translation and multiple target languages enhances compositionality modelling, and also that there is strong complementarity between distributional similarity approach and string similarity approach.

2. REFERENCES
Semi-supervised Multilingual POS tagging

[Extended Abstract]

Long Duong
Computing and Information System Department
University of Melbourne
lduong@student.unimelb.edu.au

ABSTRACT
In this paper we address the problem of multilingual part-of-speech tagging for resource-poor languages. We use parallel data to transfer part-of-speech information from resource-rich to resource-poor languages. Additionally, we use a small amount of annotated data to learn to “correct” errors from projected approach such as tagset mismatch between languages, achieving state-of-the-art performance (91.3%) across 8 languages. Our approach is based on a modest data requirements, and uses minimum divergence technique. For situations where no universal tagset mapping is available, we propose an alternate method, resulting in state-of-the-art 87.2% accuracy on the resource-poor language Malagasy.

Categories and Subject Descriptors
I.2.7 [Natural Language Processing]: Discourse

Keywords
Semi-supervised, two-layer maxent model, POS tagging, resource poor NLP, minimum divergence

1. INTRODUCTION
Part-of-speech (POS) tagging is a crucial task for natural language processing (NLP) tasks, providing basic information about syntax. Supervised POS tagging has achieved great success, reaching as high as 95% accuracy for many languages. However, supervised techniques need manually annotated data, and this is either lacking or limited in most resource-poor languages. Fully unsupervised POS tagging is not yet useful in practice due to low accuracy [1]. In this paper, we propose a semi-supervised method to narrow the gap between supervised and unsupervised approaches. We demonstrate that even a small amount of supervised data leads to substantial improvement.

Our method is motivated by the availability of parallel data. Thanks to the development of multilingual documents from government projects, book translations, multilingual websites, and so forth, parallel data between resource-rich and resource-poor languages is relatively easy to acquire. This parallel data provides the bridge that permits us to transfer POS information from a resource-rich to a resource-poor language.

Systems that make use of cross-lingual tag projection typically face several issues, including mismatches between the tagsets used for the languages, artifacts from noisy alignments and cross-lingual syntactic divergence. Our approach compensates for these issues by training on a small amount of annotated data on the target side, demonstrating that only 1k tokens of annotated data is sufficient to improve performance.

2. METHODS AND RESULTS
We first tag the resource-rich language using a supervised POS tagger. We then project POS tags from the resource-rich language to the resource-poor language using parallel word alignments. The projected labels are noisy, and so we use various heuristics to select only “good” training examples. We train the model in two stages. First, we build a maximum entropy classifier $T$ on the (noisy) projected data. Next, we train a supervised classifier $P$ on a small amount of annotated data (1,000 tokens) in the target language, using a minimum divergence technique to incorporate the first model, $T$.

Compared with the state of the art [3], we make more realistic assumptions (e.g. relying on a tiny amount of annotated data rather than a huge crowd-sourced dictionary) and use less parallel data, yet achieve a better overall result. We achieved 91.3% average accuracy across 8 languages, exceeding [3]’s result of 88.8% (error reduction ~ 22%).

The test data we employ makes use of mappings from language-specific POS tag inventories to a universal tagset. However, such a mapping might not be available for resource-poor languages. Therefore, we also propose a variant of our method which removes the need for identical tagsets between the projection model $T$ and the correction model $P$, based on a two-layer maximum entropy model over the tag pairs. Evaluating on the resource-poor language Malagasy, we achieved 87.2% accuracy, exceeding the state-of-the-art [2] of 81.2% (error reduction ~ 32%).

3. REFERENCES
One-Way Games

[Extended Abstract]

Andres Abeliuk
NICTA and University of Melbourne
andabelk@gmail.com

Gerardo Berbeglia
NICTA and Melbourne Business School
g.berbeglia@mbs.edu

Pascal Van Hentenryck
NICTA and Australian National University
pvh@nicta.com.au

ABSTRACT
This paper takes a first step into finding ways of achieving a better social welfare outcome in applications such as large-scale power restoration.

Categories and Subject Descriptors
I.2.11 [Distributed Artificial Intelligence]: Coherence and coordination

Keywords
Mechanism design, price of anarchy, social welfare

1. INTRODUCTION
This paper is motivated by optimization applications involving multiple stakeholders. Consider, for instance, the restoration of the power system and the telecommunication network after a major disaster. There are one-way dependencies between the power system and the telecommunication network. This means, for instance, that some power lines must be restored before some part of the telecommunication network can become available. It is possible to use centralized mechanisms for restoring the system as a whole. However, in practice, it is often the case that these restorations are performed by different agencies with independent objectives and selfish behavior may have a strong impact on the social welfare. It is thus important to study whether it is possible to find high-quality outcome to these problems in decentralized settings when the stakeholders proceed independently and do not share complete information about their costs.

2. ONE-WAY GAMES
A one-way game features 2 players A and B. Each player $i \in A, B$ has a public strategy set $S_i$. Each player also has a private payoff function $u_i : S \rightarrow \mathbb{R}^1$. The payoff $u_A(s_A, s_B) = u_A(s_A)$ of player $A$ is determined only by its own strategy, whereas the payoff $u_B(s_B, s_A)$ for player $B$ is determined by the strategies selected by both players. Notice that the payoff value $u_B(s_B, s_A)$ for player $B$ is known for any strategy $s_B, s_A \in S$. Player $B$ however has no information about the strategy that player $A$ will select.

Lemma 2.1. There always exists a pure Nash equilibrium in one-way games. Furthermore, Player $A$ has its optimal payoff in every Nash Equilibrium.

Our motivating applications aim at optimizing a global welfare function $SW(s_A, s_B) = u_A(s_A) + u_B(s_A, s_B)$. We quantify the quality of the worst case Nash equilibria with the price of anarchy (PoA).

Lemma 2.2. The price of anarchy for one-way games is

$$\frac{\max_{s_A \in S} u_B(s)}{\max_{s_A \in S} u_A(s)} \leq \text{PoA} \leq 1 + \frac{\max_{s_A \in S} u_B(s)}{\max_{s_A \in S} u_A(s)}.$$  

The price of anarchy can thus be arbitrarily large.

3. SINGLE-OFFER MECHANISM
We now consider a bargaining game under a Bayesian setting where each player has private utilities and a belief about the other player utilities. We assume that the default strategy $s^*_A \in \arg \max_{s_A \in S_A} u_A(s)$ of player $A$, is publicly known. This single-offer mechanism is defined as follows:

1. Player $B$ determines the strategy $s^*_A \in S_A$ and $s^*_B(s^*_A)$ that maximizes her utility.
2. Player $B$ computes the best response to player $A$’s default strategy $s^*_B(s^*_A)$.
3. Player $B$ proposes a monetary value of $\gamma \cdot \Delta_B$ with $\Delta_B = u_B(s^*_B(s^*_A)) - u_B(s^*_B(s^*_A))$ and $\gamma \in [0, 1]$ to player $A$ if she accepts to play strategy $s^*_A$ rather than her default strategy $s^*_A$.
4. If player $A$ accepts the offer, the game is played with strategy $(s^*_A, s^*_B)$; Otherwise the outcome of the game is $(s^*_A, s^*_B(s^*_A))$.

Let $F(\cdot) : F(x) \in [0, 1]$ is the probability that $\Delta_A \leq x$ and thus $F(\gamma \cdot \Delta_B)$ is the probability that player $A$ accepts the offer $\gamma \cdot \Delta_B$. $F(x)$ is assumed to be public.

We have designed a mechanism that satisfies individual rationality by construction. Player $B$ never offers more than $\Delta_B$ and its payoff is never worse than the default strategy $s^*_B(s^*_A)$.

Lemma 3.1. Player’s $B$ expected utility is maximized when she offers $\gamma^* \cdot \Delta_B$, where $\gamma^* = \arg \max_{\gamma} F(\gamma \cdot \Delta_B) \cdot (1 - \gamma)$.

We now derive the induced price of anarchy for the single-offer mechanism.

Theorem 3.2. The Bayes-Nash price of anarchy is

$$\frac{\gamma^* + 1}{\gamma^*} \cdot (1 - F(\gamma^* \cdot \Delta_B)(1 - \gamma^*)) .$$

For example, if $F = U(0, \Delta_B)$, then $\gamma^* = \frac{1}{2}$ and thus the expected price of anarchy is 2.25.
Improved $k$–centre Clustering with MapReduce

[Extended Abstract]

Jessica McClintock
University of Melbourne
j.mcclintock@unimelb.edu.au

ABSTRACT
We present an intuitive MapReduce scheme for the $k$-centre clustering problem that requires few iterations and gives a 4-approximation guarantee under certain conditions. Experimentally, we show that solutions obtained by this algorithm are comparable to a sequential 2-approximation algorithm.

Categories and Subject Descriptors
I.5.3 [Pattern Recognition]: Clustering—Algorithms

Keywords
MapReduce, clustering

1. INTRODUCTION
Clustering problems appear in fields such as social networking and bioinformatics in which the data sets can be prohibitively large. Many clustering problems are NP-hard, so it is more practical to use polynomial-time approximation algorithms which ensure solutions are within some factor of optimality. However, even when obtaining approximate solutions, there may be insufficient RAM to perform the necessary calculations. MapReduce is a scheme for parallel processing of partitions of the data, overcoming both the time and memory restrictions that arise with large data sets. The MapReduce paradigm involves several rounds of parallel and sequential computations to speed up the processing of large problems.

Definition 1. The $k$-centre problem aims to find a set of at most $k$ vertices such that the maximum distance from a vertex to its assigned centre is minimised.

Focusing on the $k$-centre problem, we consider a simple and flexible method for finding clusters with MapReduce. Given a feasible number of machines and capacity relative to the size of the data, this gives a 4-approximation for the $k$-centre problem, and only requires two MapReduce rounds.

In practice, the solutions given by our algorithm are comparable to both a sequential algorithm and an alternative MapReduce implementation, while requiring fewer MapReduce rounds and avoiding many of the restrictions of the previous MapReduce implementation [1].

2. ALGORITHMS
Gonzalez gave a simple 2-approximation algorithm for the $k$-centre problem [2]. The algorithm chooses an arbitrary vertex and marks it as a centre. At each following step, the vertex farthest from its nearest centre is marked as a new centre, until $k$ centres have been chosen.

Ene et al. [1] developed a framework for adapting $k$-centre algorithms for MapReduce. The data is sampled such that most points are well represented, and a sequential $k$-centre algorithm is run on the set of sample points. This algorithm gives a $5\alpha$-approximation with high probability, where $\alpha$ is the approximation ratio of the sequential algorithm used on the sample. Therefore, when implemented using the 2-approximation algorithm described above, with high probability this results in a 10-approximation.

Given $m$ machines, each with capacity $c$, and input set $V$ containing $n$ points, we present the following method of adapting $k$-centre algorithms to the MapReduce framework.

Algorithm 1 MapReduce–$k$-centre ($V$, $k$, $m$)
1: Arbitrarily partition $V$ into sets $V_1,\ldots,V_m$ such that $\bigcup_i V_i = V$ and $|V_i| \leq \lceil n/m \rceil$.
2: Send each set $V_i$ to a machine $\rho_i$.
3: In parallel, each machine $\rho_i$ runs a $k$-centre algorithm on $V_i$, and returns $S_i$, the $k$ centres found.
4: Send $S = \bigcup_i S_i$ to a single machine.
5: Run a $k$-centre algorithm on $S$; return the centres $S^*$.

Theorem 1. If $n \leq m \cdot c$ and $k \cdot m \leq c$, then the $k$-centre algorithm can be implemented in MapReduce with a 4-approximation guarantee.

This results from implementing our algorithm with the 2-approximation algorithm of Gonzalez [2] as a subprocedure.

Preliminary experimental results demonstrate that both MapReduce algorithms generally perform as well as the 2-approximation algorithm of Gonzalez. Ene’s algorithm appears to take many rounds, and give poor solutions, when $k$ is very small and the number of dimensions in the data is large. The most significant factor in the runtime of MapReduce algorithms on dense graphs is sending the data between machines, so it is important to use few rounds. Our algorithm is designed to only use two passes under most conditions; the sampling algorithm uses at least four. For larger $k$, the Ene algorithm is theoretically equivalent to the Gonzalez algorithm, as shown in Table 1 for $k = 25$.

3. CONCLUSIONS
Our algorithm is simpler, requires fewer MapReduce rounds, and is able to find larger numbers of clusters than the alternative MapReduce algorithm of Ene et al. [1].

4. REFERENCES
Designing Organization-Aware Agents using OJAzziIC

[Extended Abstract]

Kathleen Keogh
PhD Student, Computing and Information Systems Dept. University of Melbourne
Academic, School of Engineering and Information Technology,
Federation University Australia, Ballarat
k.keogh@federation.edu.au

ABSTRACT

In this paper, we highlight design considerations when using the OJAzziIC architecture in multi-agent system design. We list a number of steps to be considered in the specification of agent requirements. We have used these in the design and creation of an agent based incident response simulation system.

Categories and Subject Descriptors
I.2.11 [Artificial Intelligence]: Distributed Artificial Intelligence—Intelligent agents

Keywords
Multiagent systems, coordination, adaption, organizations

1. INTRODUCTION

We are interested in the specification and design of flexible and adaptive agent organizations. OJAzziIC [1, 2] defines an architecture for agent organizations using contracts to coordinate plans and beliefs. These contracts define agreed organizational plans and explicit social policies regulating communication and knowledge sharing within the organization. Such organizationally aware agents are suited to dynamic domains such as incident response in emergency management. In such complex environments, agents cannot rely on pre-scripted plans for coordinated behaviour, but must dynamically coordinate knowledge and plans.

The OJAzziIC model, in addition to contracts, includes a role model, a hierarchical goal tree and an organizational plan [1]. Social policies also make explicit the agents’ obligations within an organization in terms of role adoption, selection of goal objectives and communication [2]. Identification of these components is important in the design phase and will be addressed in this paper.

2. DESIGN CONSIDERATIONS

When a problem is specified, some details can be described at design time. Long term organizations can be identified based on agencies in the real world such as the medical response organization (Ambulance Agency) and the law enforcement organization (Police Force). Within these organizations, roles and responsibilities can be specified a priori. During an actual crisis management scenario, new additional, informal organizations emerge dynamically and can cross existing organizational boundaries. These organizations persist over some time to assist with coordination of particular objectives and to facilitate inter-organizational coordination. The driving forces for the creation of a dynamic adhocracy (adhoc organization [2]) include the needs for coordinated behaviour and coordinated knowledge sharing. Autonomy and initiative allowed in agents’ organizational reasoning needs to be determined. For example, the task of clearing away bystanders may be fulfilled by any agent (Officer or Medic) within the vicinity using their initiative without being specifically asked/assigned. However, the task of rescuing an injured agent and moving them to the ambulance might only be adopted by a Medic agent as allocated by the Medic in charge.

In order to create a dynamic OJAzziIC organization, at design time, the following steps should be addressed:

- Create a high level goal decomposition of objectives and where possible, break objectives into tasks.
- Identify dependencies between tasks and objectives, paying attention to requirements of synchronisation – e.g. before(task1,task2), concurrent(task1,task2).
- Identify capabilities required to achieve each task.
- Identify long term organizations agents may belong to.
- Identify autonomy and control associated with each objective or task.
- Identify roles that agents may be able to adopt within each organization (may include domain roles and structural roles).
- Identify responsibilities associated with roles
- Identify role relationships (e.g. dependency, authority, right to delegate etc.).
- Establish social policies to be adopted within the organizational contract:
  - role adoption responsibilities.
  - knowledge sharing obligations.
  - organizational adhocracy creation triggers
  - obligations between agents to establish shared organizational plans for coordinated tasks.

We have implemented a simple simulation incident response system using this process in order to clarify the design methodology.

3. REFERENCES

Reliable power transmission networks

[Extended Abstract]

Rodrigo R. Gumucio E.
The University of Melbourne & NICTA
RGumucio@student.unimelb.edu.au

Carleton Coffrin
NICTA & The University of Melbourne
Carleton.Coffrin@nicta.com.au

Pascal Van Hentenryck
The Australian National University & NICTA
Pascal.VanHentenryck@anu.edu.au

ABSTRACT
In the context of the Transmission Network Expansion Planning (TNEP) problem, a recent study shows that there exists a gap, between high cost expansion plans produced by AC-heuristics and potentially infeasible expansion plans produced by DC-approximations, that can be bridged by using the recently proposed LPAC power flow approximation.

The work in progress for this paper aims to incorporate the $n-1$ reliability criterion to the mentioned expansion planning study by proposing a genetic algorithm which may scale to large networks.

Categories and Subject Descriptors
I.2 [Artificial Intelligence]: Miscellaneous;
G.1.6 [Optimisation]: Constrained optimisation

Keywords
Power transmission network planning, $n-1$ reliability, linear ac power flow, non-linear non-convex optimisation, genetic algorithm, ac feasible expansion plans

1. INTRODUCTION

Power transmission networks need to be expanded, for example, to satisfy a projected future demand or to make them more robust against potential threats like natural disasters. Expanding a network entails to add a set of new circuits to it which is typically very expensive. We strive then to choose a set of circuits, called an expansion plan, such that the investment cost of adding them to the network is minimum.

The minimisation problem is hard because, due to the nature of the AC power flow model, it is subject to complex non-convex and non-linear constraints which not only grow with the number of nodes in the network but also involve, among a huge number of model variables, both discrete and continuous decision variables.

The intractability of this problem is, in most studies, worked around by using a heuristic to the full AC model or by relying on an approximation based on the popular DC linear power flow model. However, due to the limitations of DC-approximations which ignore reactive power and voltage magnitudes, recent work in the community started to consider the TNEP problem with the full AC power flow equations (AC-TNEP), as for example [2].

We alternatively take the new LPAC[3] power flow model, which attempts to provide a good balance between the tractability of the DC model and the accuracy of the AC model, and propose a LPAC-TNEP formulation which proves to be a good mechanism to find expansion plans of reasonable cost with no network violations.

<table>
<thead>
<tr>
<th>Case</th>
<th>MIP without $n-1$</th>
<th>GA with $n-1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost (#) Violations</td>
<td>Cost (#) Violations</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>130 (5) 0.00</td>
<td>200 (8) 0.00</td>
</tr>
<tr>
<td>24</td>
<td>681 (15) 0.00</td>
<td>1557 (36) 0.00</td>
</tr>
</tbody>
</table>

2. PRIOR WORK AND RESULTS
The Transmission Network Expansion Planning problem asks to find a lowest cost expansion plan to satisfy a projected future energy demand. Our study in [1] found that an LPAC-TNEP formulation, which can benefit of industrial Mixed Integer Program (MIP) solvers, can be used to find AC feasible expansion plans of reasonable cost for several benchmarks. For illustrative purposes, results are shown in the left side of Table 1 for the 6 and 24 bus benchmarks. Evidence shows, however, that even using state-of-the-art industry standard tools, the MIP will struggle and even fail to find feasible solutions for larger benchmarks or even for the same benchmarks used in [1] when the reliability constraints are included into the model.

3. WORK IN PROGRESS AND RESULTS
When deciding which circuits to add to an expansion plan, usually a reliability criteria is desired to be taken into account in the analysis in order to make the network more robust. We are studying now a genetic algorithm (GA) for optimisation which relies on an LPAC-based simulator to check for $n-1$ feasibility. The main characteristics of the GA are a fitness function which drives towards both feasibility and optimality; a two point crossover operation; and a mutation operation with both an increasing mutation probability and a filter based on the metropolis probability.

Current results, as shown in the right side of Table 1, give some evidence that it may be possible to find $n-1$ compliant AC feasible expansion plans of reasonable low cost with an acceptable number of not so expensive simulations.

4. REFERENCES
ABSTRACT
In this paper, we describe how Neighborhood Component Analysis can be modified as the feature learning step for learning from high dimensional and class imbalanced data.

Categories and Subject Descriptors
I.2.6 [Artificial Intelligence]: Learning; I.5.4 [Pattern Recognition]: Applications; I.5.3 [Pattern Recognition]: Clustering – similarity measures; I.5.2 [Pattern Recognition]: Design Methodology – Feature evaluation and selection

Keywords
NCA, feature learning, distance measure, class imbalance

1. INTRODUCTION
Feature learning from high dimensional small sample data is difficult. Deep learning models can be used for feature learning from data. However, if the dataset is not large enough, there is some evidence suggesting that these models are prone to overfitting. Class imbalance is another problem that feature learning models need to address.

Neighborhood Component Analysis (NCA) [1] is a metric learning method that learns the Mahalanobis distance measure between samples. Nonlinear extension of NCA [2] learns the distance measure on the feature space learned by unsupervised pre-training instead of input feature space.

In this paper, we show that NCA can learn the features from high dimensional data even when the dataset is small. However, it does not consider the class imbalance problem and as the result it can overfit to the majority class. We modify the NCA objective function to address class imbalance problem.

2. APPROACH
We use the nonlinear NCA as our model because it can be extended it into a deep learning model. However, the experiments presented in this paper are conducted on a single layer shallow model. Restricted Boltzmann Machine (RBM) [3] is used to learn the features by unsupervised training and initialized the values of the weight matrix. NCA or class imbalance normalized NCA (CIN-NCA) is used for fine-tuning the weights. CIN-NCA is the NCA with the modified objective function to balance the effect of sample size in each class and to consider the dissimilarity between samples.

NCA assumes that the Euclidean distance in learned feature space is the same as Mahalanobis distance in input feature space. NCA objective function projects the similar samples in the original sample space to be closer in the transformed space. The similarity is defined based on the class labels.

3. RESULTS
For learning features from high dimensional data with small sample size, we sample 3000 images of hand-written digits from MNIST dataset. Each image has 784 pixels. 1000 images are used for training, validation and testing each. The comparison of the accuracy of k-nearest neighbor classifier and SVM, with or without feature learning, is shown in Table 1. Classification after feature learning step improves the accuracy of classifiers. When we trained the Deep Belief Net (DBN) [3] with 1 hidden layer of 500 nodes without NCA, we obtained the test error of 10.8% and this error is similar to the test error of SVM. The test error of the k-nearest neighbor is higher without feature learning. However, as NCA is optimizing the assumption of nearest neighbor classifier, it achieved better performance after learning features with NCA.

Table 1. Test error on MNIST 1000 samples

<table>
<thead>
<tr>
<th>Feature Learning</th>
<th>KNN (k=5)</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>12.7</td>
<td>10.6</td>
</tr>
<tr>
<td>NCA</td>
<td>9.2</td>
<td>9.5</td>
</tr>
<tr>
<td>CIN-NCA</td>
<td>6.9</td>
<td>9.3</td>
</tr>
</tbody>
</table>

The results of experiment for class imbalance problem are described in Table 2. We used E. coli dataset which is a highly imbalanced dataset for binary classification. The proportion of the classes is 35:301. The data has 7 dimensions. 235 samples are used for training, 71 for validation, and 101 for testing. The first F1-score is for pre-training with Bernoulli RBM and the second (in parentheses) is for Gaussian-Bernoulli RBM. The results suggest that our model can handle the class imbalance problem.

Table 2. F1-score on E. coli dataset

<table>
<thead>
<tr>
<th>Class</th>
<th>Support</th>
<th>NCA F1-score</th>
<th>CIN-NCA F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>87</td>
<td>0.93(0.96)</td>
<td>0.96(0.97)</td>
</tr>
<tr>
<td>+1</td>
<td>14</td>
<td>0.24(0.70)</td>
<td>0.67(0.77)</td>
</tr>
</tbody>
</table>

4. REFERENCES
Mining User Interactions and Activities using Mobile Data while Maintaining Privacy

[Extended Abstract]

Anthony Quattrone
Department of Computing and Information Systems, The University of Melbourne
quattronea@unimelb.edu.au

ABSTRACT
Mobile privacy has recently become a well-researched topic in pervasive computing. However more research is needed to determine if personal information can be inferred from raw mobile data. Examples include if users are meeting other users, calls a user is making and receiving, trajectories of individuals accessing location-based services, social media services used and if mobile sensors can be used to identify a person. We demonstrate sophisticated algorithms that reveal a user’s interactions and activities based on data stored on a mobile device. Mobile users typically keep their devices with them at all times. Near-field technologies (Bluetooth, WiFi) can be used as sensors to determine if user devices are in proximity to each other. Through the collection of very granular mobile data, we demonstrate that it be inferred when users are meeting one another.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data Mining

Keywords
Mobile privacy, predictive modeling, inference attacks

1. INTRODUCTION
Mobile smartphones have become ubiquitous. The success of these devices combined with advancements in mobile network infrastructure have created a strong market for third party apps, cloud storage providers and location based service (LBS) providers.
Smartphones capture a significant amount of personal and sensitive information about a user. We hypothesise that data internal to these devices sent via a network can be combined with rich external data to recreate a very accurate profile about an individual.
Modern mobile platforms allow for easy collection of user data via public APIs. Data can be sent to a remote server without the user being aware. Mobile applications are requesting excessive use of permissions in the Google Play store relative to the advertised functionality they provide. Raw data from mobile sensors can be easily and stored in a relational database. While it is often speculated that personal information can be inferred from sensitive data, literature to highlight tools and techniques utilised to invade user privacy is scarce. We aim to to use raw data from a variety of sensors on a mobile device to determine user behaviours by aggregating the data and deriving variables. With the

<table>
<thead>
<tr>
<th>Train/Test Split (%/%)</th>
<th>Classification Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70/30</td>
<td>98.57</td>
</tr>
</tbody>
</table>

ability to easily collect mobile data via third party apps using only standard APIs, a high speed Internet connection that can store data in a remote database and applying advanced machine learning techniques can put a smartphone user’s privacy in jeopardy.

The focus of this research is to establish if it can be determined if people are actually meeting one another in social interactions.

The key contributions of our research are as follows:

- Providing a system to capture mobile data in a structured format to allow for analysis integration
- Investigate the privacy implications of mobile data collection
- Determine if patterns can be mined from mobile data to determine if users are meeting with other users
- Provide a framework to protect users from pattern mining attacks

1.1 Preliminary Results
Using only diagnostic features (e.g. User Settings), Naive Bayes is accurate in predicting a user.

Our sample for analysis contained 223 days of data in which 66 devices could be uniquely identified. In the preliminary analysis we trained and tested our model on devices that contained at least two days of captured data, however we found that this was not sufficient to uniquely determine a user. Thus, only devices in the data set with at least three days of captured data were considered for analysis. In practice three days is still a small amount of time as user’s will use apps for significantly longer periods which is likely to increase the accuracy of our approach further.

The dataset was distributed ensuring that each device has 70% of the data points as training data and evaluated on the remaining 30%.
ABSTRACT

In collaborative anomaly detection, multiple data sources submit their data to an on-line service, in order to detect anomalies with respect to the wider population. A major challenge is how to achieve reasonable detection accuracy without disclosing the actual values of the participants’ data. We propose a lightweight and scalable privacy-preserving collaborative anomaly detection scheme called RMP, which is a combination of nonlinear and participant-specific linear perturbation. A privacy analysis is given for Bayesian Estimation and ICA attacks. Experimental results on various datasets using an autoencoder show that RMP yields comparable results to non-privacy preserving anomaly detection.

Categories and Subject Descriptors
K.6 [Information management]: Security & protection

Keywords
Privacy-preserving data mining, anomaly detection.

1. INTRODUCTION

Anomaly detection plays a key role in data mining for detecting unusual patterns or events in an unsupervised manner. In particular, there is growing interest in collaborative anomaly detection [1], where multiple data sources submit their data to an on-line service, in order to detect anomalies with respect to the wider population. For example, in participatory sensing networks (PSNs), participants collect and upload their data to a central service to detect unusual events, such as the emergence of a source of pollution in environmental sensing, or disease outbreaks in public health monitoring. A major challenge for collaborative anomaly detection in this context is how to maintain the trust of participants in terms of both the accuracy of the anomaly detection service as well as the privacy of the participants’ data. In this paper, we propose a random perturbation scheme for privacy-preserving anomaly detection, which is resistant to a variety of attacks while achieving comparable accuracy to anomaly detection on the original unperturbed data.

There have been several studies on collaborative anomaly detection, where a number of participants want to build a global model from their local records, while none of the participants are willing to disclose their private data. Most existing works are based on SMC. Achieving high levels of privacy and accuracy, SMC incurs a high communication and computational overhead. Moreover, SMC based methods require the coordination of all participants during the training process, limiting the number of participants. Thus, an open research challenge is how to improve scalability while achieving high levels of accuracy and privacy.

To address this challenge, we propose a privacy-preserving scheme for anomaly detection called Random Multiparty Perturbation (RMP). RMP supports the scenario where participants contribute their local data to a public service that trains an anomaly detection model from the combined data. This model can then be distributed to users who want to test for anomalies in their local data.

In order for the participants of RMP to maintain the privacy of their data, we propose a form of random perturbation to be used by each participant. Previous approaches to random perturbation, e.g., [1, 2], require all participants to perturb their data in the same way, which makes this scheme potentially vulnerable to breaches of privacy if collusion occurs. While, in RMP each participant first perturbs their data using a unique, private random perturbation matrix. Moreover, any user can apply the resulting anomaly detection model to their own local data by using a public perturbation matrix. This provides a scalable collaborative approach to anomaly detection, which ensures a high level of privacy while still achieving a high level of accuracy.

The main contributions of this paper are as follows: i) We propose a privacy-preserving model of collaborative anomaly detection based on random perturbation, such that each participant in training the anomaly detector use their own unique perturbation matrix, while the resulting anomaly detection model can be used by any number of users for testing. ii) We give the first privacy-preserving scheme for autoencoders, that does not rely on computationally intensive cryptographic techniques. iii) We show analytically the resilience of RMP to Bayesian Estimation and ICA attacks. iv) We show that the accuracy of RMP is comparable to non-privacy preserving anomaly detection on various datasets.

2. REFERENCES

Automatically Recognizing Places of Interests/Activities from Unreliable GPS Data and Smart Phone Sensors

[Extended Abstract]

Tanusri Bhattacharya
Department of Computing and Information Systems
The University of Melbourne
Parkville, Australia
tbhattachary@unimelb.edu.au

Lars Kulik
Department of Computing and Information Systems
The University of Melbourne
Parkville, Australia
lkulik@unimelb.edu.au

James Bailey
Department of Computing and Information Systems
The University of Melbourne
Parkville, Australia
baileyj@unimelb.edu.au

ABSTRACT

Modern smartphones are equipped with many sensors, such as GPS, Wi-Fi, BlueTooth, ambient temperature, pressure, humidity sensors, etc. This has created opportunity to the researchers in pervasive community to mine useful semantic context about a user’s behaviour. One important user’s context is to mine the significant places visited by a user. GPS provides broad coverage for positioning worldwide. It provides excellent accuracy at outdoor environments and does not incur any infrastructural cost for indoor positioning. Significant places are often indoor locations and may correspond to buildings in urban areas where GPS is unreliable. In this research, we develop a novel algorithm POI-ID to identify a user’s Place of interests (POIs) at building level accuracy from highly inaccurate GPS data. We further extend our research to explore the ambient sensors embedded in a smartphone to recognize a user’s places or activities in real time.

Categories and Subject Descriptors

H.2.8 [Database Applications]: Spatial databases and GIS

Keywords

Indoor GPS, urban canyons, trajectory mining, significant places, smartphone sensors

1. INTRODUCTION

The rise of sensor equipped (GPS, Wi-Fi, etc.) smartphones has created exciting opportunities to mine useful knowledge about users’ moving behaviour: for example, the places visited, the activities performed, etc. This knowledge can be used in a range of application purposes: for example health monitoring, offender monitoring, traffic and transportation management systems, location-based services and social networks.

Although satellite-based GPS positioning provides good coverage for modern navigation and tracking; it does not work well for indoor places or urban canyons, due to poor line-of-sight transmission between the receiver and the satellites. The complexity of indoor environments causes the multipath propagation of a GPS signal and severely reduces the accuracy of positioning. In urban canyons, the multipath effect and signal loss are also caused by surrounding tall buildings. A range of prior approaches proposed the use of other technologies such as RFID, Wi-Fi or Bluetooth to improve location accuracy [2, 3, 1]. However, the opportunity to use such technologies is dependent on the building-infrastructure and may not be always available. Typically, there is also additional cost for implementation.

Indoor places play an important role in daily life and people spend a large amount of their time in indoor locations. GPS provides broad coverage for positioning worldwide. It provides excellent accuracy for outdoor environments and does not incur any additional infrastructure cost for indoor positioning. However, a user’s Place of Interests (POIs) are often indoor locations and may correspond to buildings in urban areas where GPS is less accurate. In this research, we develop a novel algorithm (POI-ID) to obtain a user’s POI(s) from a set of highly inaccurate GPS data points. Given a set of time-stamped GPS points and a list of surrounding POIs, POI-ID produces a ranking of the POIs according to the user’s most likely location at a building level accuracy. An experimental study demonstrates the superiority of our algorithm against several baseline approaches with a recall of 96.5% for the top 5 retrieved locations.

POI-ID can successfully predict a user’s POI at building level accuracy. However, for situations like fire-fighting and search and rescue operations, we may need to track users in real-time at a room level accuracy. Furthermore, the major challenge of using GPS-enabled mobile devices at indoor places is the large power consumption. We have extended our research to investigate some of these challenges by exploring the ambient sensors embedded in a smartphone.

2. REFERENCES


Principled Dictionary Pruning for Low-Memory Corpus Compression

[Extended Abstract]

Jiancong Tong
Department of Computing and Information Systems,
The University of Melbourne, Australia
lingfenghx@gmail.com

ABSTRACT

Compression of collections, such as text databases, can both reduce space consumption and increase retrieval efficiency, through better caching and better exploitation of the memory hierarchy. A promising technique is relative Lempel-Ziv coding, in which a sample of material from the collection serves as a static dictionary; in previous work, this method demonstrated good compression ratios, while allowing extremely fast random access to individual items. However, there is a trade-off between dictionary size and compression ratio, motivating the search for a compact, yet similarly effective, dictionary. In previous work it was observed that, since the dictionary is generated by sampling, some of it (selected substrings) may be discarded with little loss in compression. Unfortunately, simple dictionary pruning approaches are ineffective. We generate measures for identification of low-value substrings in the dictionary, and show on a variety of sizes of text collection that halving the dictionary size leads to only marginal loss in compression ratio. This is a dramatic improvement on previous approaches.

Categories and Subject Descriptors
E.4 [Coding and Information Theory]: Data compaction and compression

Keywords
Corpus compression, string algorithms, optimization

1. INTRODUCTION

For web-scale collections, relative Lempel-Ziv factorization (RLZ) [1] is a good compression scheme that achieves both good compression effectiveness and high decompression speed. With RLZ, the collection text is parsed into a contiguous sequence of fragments, where each fragment is sourced from an external static dictionary. In the dictionary-generation method (GEN) originally proposed by Hoobin et al. [1], fixed-size blocks of data (say 1 KB) are sampled from the repository and then concatenated to form the dictionary.

However, it is possible that some strings are sampled many times (as would be expected, statistically) during the dictionary generation procedure, meaning that there is extensive redundancy in the dictionary and it is larger than required. Hoobin et al. [2] observed that some parts of the dictionary were rarely, or even never, used. Thus, they proposed an elimination method (we refer it as REM and use it as our baseline) which removes the least frequently referred to blocks. RLZ with algorithm REM compresses better, and is faster, than LZMA when the dictionary size is halved, while it outperforms ZLIB even with a ten-fold reduction in dictionary size. However, our preliminary experimental results suggest that there are striking differences between the reference frequency of various parts of a sample block. Thus it is far from optimal to treat the block as the unit of elimination. Therefore, pruning at a finer granularity is needed.

A major challenge in dictionary pruning is how to choose the segments to remove. In this work, instead of estimating the consequence of removing a segment, we propose a heuristic contribution-aware reduction method (CARE) which estimates the ‘contribution’ of a segment (to compression effectiveness) if it is kept in the pruned dictionary.

We generate measures for identification of low-value substrings in the dictionary. The measures can be calculated efficiently by factoring segments of the dictionary against the dictionary itself. By eliminating low-value segments, we can markedly reduce the volume of the dictionary without significant loss of compression performance. As shown in Table 1, on a 426 GB collection (GOV2) and a dictionary of 1000 MB, the data is reduced to 10.271% of its original size; halving the 1000 MB dictionary to 500 MB increased compressed size by 0.276% (in absolute terms) with REM, whereas with CARE it increases by only 0.005%. Halving again to 250 MB gives increases of 3.181% (previous method) and 0.636% (our method), respectively. Note that IDS/PDS in the table stand for initial/pruned dictionary size.

<table>
<thead>
<tr>
<th>PDS (MB)</th>
<th>IDS = 2000 MB</th>
<th></th>
<th>IDS = 1000 MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEN</td>
<td>REM</td>
<td>CARE</td>
<td>GEN</td>
</tr>
<tr>
<td>1000</td>
<td>10.271</td>
<td>9.588</td>
<td>9.437</td>
</tr>
<tr>
<td>250</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

2. ACKNOWLEDGMENTS

I would like to thank my supervisors Justin Zobel and Anthony Wirth for their guidance throughout this work.

3. REFERENCES


Predicting at-risk students in Massive Open Online Courses

[Extended Abstract]

Jiazhen He
Department of Computing and Information Systems
The University of Melbourne
jiazhenh@student.unimelb.edu.au

James Bailey
Department of Computing and Information Systems
The University of Melbourne
baileyj@unimelb.edu.au

Rui Zhang
Department of Computing and Information Systems
The University of Melbourne
rui.zhang@unimelb.edu.au

Benjamin Rubinstein
Department of Computing and Information Systems
The University of Melbourne
benjamin.rubinstein@unimelb.edu.au

ABSTRACT
Massive Open Online Courses (MOOCs) have received wide public attention for their potential to scale higher education with multiple platforms such as Coursera, edX and Udacity. However, a big problem MOOCs face is the high rates of attrition. We explore to predict at-risk students early and accurately, and provide possible interventions for them.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data Mining

Keywords
MOOC, learning analytics, educational data mining, prediction model, logistic regression

1. INTRODUCTION
With the booming popularity of Massive Open Online Courses (MOOCs), such as Coursera, edX and Udacity, MOOCs have attracted growing attention from educators, computer scientists and general public in recent years. MOOCs aim to make higher education accessible to entire world, by offering various online courses from universities for free, and has attracted a diverse population of students from all age group, education background and nationalities. Nevertheless, MOOCs face a big problem, the high rates of attrition. For example, Table 1 shows the student participation on two courses from The University of Melbourne delivered onCoursera, which indicates the extremely high dropout rate.

In this paper, we focus on identifying the students who are at risk of completing the course as early as possible based on the information available by the end of each week, so as to provide them appropriate interventions. We investigate the above two courses, and attempt to address the following questions.

(a) Can we identify the students who are at risk of completing the course early and accurately?
(b) Can we provide possible interventions for students who are at risk?

Table 1: Students participation on Courses The Principle of Macroeconomics (Maroc) and Discrete Optimization (DisOpt); actions are measured in terms of viewing, downloading lectures and doing quizzes or assignments.

<table>
<thead>
<tr>
<th></th>
<th>Macro</th>
<th>DisOpt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of students enrolled</td>
<td>66,204</td>
<td>51,395</td>
</tr>
<tr>
<td>Number of students with no actions</td>
<td>39,908</td>
<td>30,044</td>
</tr>
<tr>
<td>Number of students completed</td>
<td>1,412</td>
<td>795</td>
</tr>
</tbody>
</table>

Before building prediction models, data preprocessing is required. Students enrolled for various motivations and needs. We are interested in helping those who intend to pass the course. Therefore, we focus on students who are active in doing assignments/quizzes, which might indicate their intention to pass the course. Then we extract features for classification from student engagements in video lectures and assignments/quizzes.

(a) To address the first question, we build and compare different prediction models by the end of each week, and find that logistic regression performs best in terms of accuracy and AUC. Furthermore, logistic regression is preferred due to its good interpretability, which is important for educational community.

(b) To address the second question, one possible intervention is to present students real meaningful probability of failing to help them realize their progress. In particular, we focus on intervening those who are on the borderline to pass rather than high-level of risk students since telling them might push them away the course. We use logistic regression, which can produce well-calibrated probabilities.

Another possible intervention can be obtained from decision tree. The paths from the root to the leaves can capture the characteristics of students who pass and fail, and therefore can be used for intervention.
Improved Feature Transformations for Classification using Density Estimation

[Extended Abstract]

Yamuna Kankanige  
Department of Computing and Information Systems  
University of Melbourne  
ykankanige@student.unimelb.edu.au

James Bailey  
Department of Computing and Information Systems  
University of Melbourne  
baileyj@unimelb.edu.au

ABSTRACT

Density based logistic regression (DLR) is a classification technique that has been introduced recently. It can be explained as a one-to-one non-linear transformation of the original feature space based on density estimations, which is particularly suitable for learning a logistic regression model. Performance gains, good interpretability and time efficiency make DLR an attractive technique. However, there are some open questions regarding the use of DLR. In our work, we tackle these limitations and propose several new extensions which are evaluated using several synthetic and publicly available datasets.

Categories and Subject Descriptors
H.2.8 [Database applications]: Data mining

Keywords
Classification, feature transformation, kernel density estimation

1. INTRODUCTION

Logistic regression (LR) is a popular non-linear classifier, because LR estimates the probabilities of classes directly and the coefficients of the resulting model provide information about the relative importance of input features. DLR [1], proposed by Chen et al., is a non-linear one-to-one mapping of the original features into another feature space based on density estimations. This new feature space is most appropriate for training a LR model for classification.

Feature transformation is a popular machine learning technique, which might be either supervised or unsupervised. Transforming feature \( x \) in to \( x^k, \log x, \sin x, 1/x \) or normalising into the range \([0, 1]\) are examples of unsupervised transformations, whereas supervised discretization of continuous features to discrete features using the class labels is an example of supervised feature transformations. Transformations are useful for modelling non-linearity, stabilizing variance in data, placing features on equal scales or constructing feature spaces with better separation between classes. DLR can be considered as a supervised non-linear feature transformation technique, which is constructing a feature space with better separation between classes.

Figure 1 illustrates how feature transformations in DLR work using a synthetic dataset, which compares the projections of the original and transformed values of a feature.

As we can see, transformed feature values provide better separation between classes compared to the original feature values. However, we could identify some limitations to the formulation of DLR. For example, ensuring robust density estimations is important, which has not been addressed in the original work.

In our studies, we address the following questions regarding the use of DLR.

(a) How can we provide robust density estimations?
(b) Is it beneficial to extend this to higher order transformations?
(c) What are the situations where DLR is effective?

As answers to the above questions, we propose the following ideas, which are the outcomes of our studies.

(a) A methodology that can be followed in density based transformations, where we propose that a separate dataset has to be used for density estimations.
(b) A technique that performs many-to-one transformations, based on the use of higher order kernel density estimation, and evidence that it can sometimes be more effective than the original method. However, transformations beyond two dimensions do not appear to be advantageous.
(c) Evidence that density based transformations are useful for classification in transfer learning scenarios.

2. REFERENCES

FILTA : Better View Discovery from Collections of Clusterings via Filtering

[Extended Abstract]

Yang Lei       Nguyen Xuan Vinh       Jeffrey Chan       James Bailey
Department of Computing and Information Systems
University of Melbourne
yalei@student.unimelb.edu.au
{vinh.nguyen, jeffrey.chan, baileyj}@unimelb.edu.au

ABSTRACT

Traditional clustering methods have focused on finding a single ‘best’ clustering solutions. However, in real applications, there may exist more than one reasonable set of clusters (a clustering) to explain the data. Also, users are often uncertain about what is a good clustering for their purpose. Meta clustering has been successfully used to discover multiple reasonable and different clusterings (views) in a dataset by navigating and refining a large collection of base clusterings. However, the effectiveness of meta-clustering is highly dependent on the distribution of the base clusterings and open challenges exist with regard to its stability and noise tolerance. In this paper we propose a simple and effective filtering algorithm (FILTA) that can be flexibly used in conjunction with any meta-clustering method and get better views by removing base clusterings having poor quality or high redundancy. We evaluate FILTA on both synthetic and real world datasets, and see how its use can enhance view discovery for complex scenarios.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—clustering

Keywords

Clustering, meta-clustering, multiple clusterings, clustering visualization

1. INTRODUCTION

Clustering is one of the most important unsupervised techniques for discovering dataset structure. Many clustering methods focus on obtaining one single ‘best’ solution by optimizing a pre-defined criterion. There are two limitations with this: firstly, data can be multi-faceted in nature. Particularly when the datasets are large and complex, there may be several useful clusterings that exist, not only one. Secondly, users may be seeking different perspectives on the same dataset, requiring multiple clustering solutions.

Multiple clustering analysis [1] aims to discover a set of reasonable and distinctive clustering solutions from the same dataset. Many methods have been proposed on this topic and one very popular technique is meta-clustering. Meta-clustering first generates a large number of base clusterings. These base clusterings may then be meta-clustered into groups. Further, clusterings within the same group can be abstracted to a consensus view of that group. This results in one or more distinctive clusterings (views) of the dataset, each offering a different perspective or explanation.

A major drawback and challenge with the use of meta-clustering is that its effectiveness is highly dependent on the quality and diversity of the generated base clusterings. We illustrate this problem with an example in Figure 1a, where the dataset consists of four Gaussian clusters. We generate a set of raw base clusterings with 2 clusters via different procedures. These base clusterings are then meta-clustered into groups, and for each group a consensus view is extracted via consensus clustering. The views generated on the raw base clusterings are presented in Figure 1a. Observe that among these four views, some are rather similar (view2, view3 and view4) and some have poor quality (view1, view2 and view4). Can we apply a filtering step to the base clusterings and thus avoid discovering poor quality or redundant views? Figure 1b provides intuition about the benefits of filtering. It shows the views generated after filtering the raw base clusterings, which are of high quality and non-redundant. In this paper, we propose filtered meta-clustering (FILTA), aiming at detecting multiple high quality and distinctive views by filtering and analyzing a given set of base clusterings.

2. REFERENCES

Bandwidth-Efficient Convergence of Gossip-based Recommendation Systems

Irum Fahim Bukhari, Aaron Harwood, and Shanika Karunseker
ibukhari@student.unimelb.edu.au, aharwood@unimelb.edu.au, karus@unimelb.edu.au

ABSTRACT
In this work, we analyzed the effect of similarity distributions on the convergence of the protocol and found the trade-off between convergence time and bandwidth to decide optimal settings that lead to fast convergence and affordable bandwidth from recommendation systems perspective.

Categories and Subject Descriptors
H.3.4 [Information storage and retrieval]: Distributed systems; H.3.3 [Information storage and retrieval]: Information filtering

Keywords
Semantic overlays, gossip protocols, peer-to-peer networks, collaborative filtering, recommendation systems

1. INTRODUCTION
Collaborative Filtering based recommendation systems are gaining attention of researchers due to the evolution of web 2.0 where users of Internet are more than just consumers of information. Collaborative Filtering (CF) can be implemented using gossip-based similarity overlays. Efficient recommendations in such systems would depend on how fast they converge upon user profile changes. We analyze the convergence behavior of the VICINITY protocol [1] with different types of similarity distributions, under various protocol settings. Second, we find the settings of the protocol parameters that give the minimum convergence time (approximately), from a worst case random structure but also avoid wastage in bandwidth. Optimal settings involve the trade-off between the convergence time and the bandwidth utilization. Our simulations show that there exists a threshold for gossip length that gives the best convergence speed eliminating the need of excessive bandwidth.

1.1 Experiments
In Figure 1, convergence time for distributions Mesh, Torus, Uniform, Poisson, and Powerlaw are shown. Bars represent view sizes in sequence \{4, 6, 10, 15, 20, 25, 30, 35, 40, 45\} moving from left to right for each distribution. Figure 2 shows the threshold gossip length for each view size. Figure 3 shows the bandwidth utilization for different network sizes.

2. REFERENCES
Developing an opportunistic online psychosocial screening tool for young people in general practice

[Extended Abstract]

Marianne Webb  
General Practice and Primary Health Care Academic Centre  
Melbourne Medical School  
webbm@unimelb.edu.au

Greg Wadley  
Interaction Design Lab  
Computing and Information Systems  
Melbourne School of Engineering  
greg.wadley@unimelb.edu.au

Sylvia Kauer, Lena Sanci  
General Practice and Primary Health Care Academic Centre  
Melbourne Medical School  
sylvia.kauer@unimelb.edu.au,  
l.sanci@unimelb.edu.au

ABSTRACT

Online self-administered multidimensional screening tools are a quick and effective way for General Practitioners (GPs) to identify areas of health concern for young people. However, there is currently little research into how they can be designed in a way that places young people as more active participants in their consultation. There is also a paucity of research that explores the design needs of GPs and how to successfully integrate these tools into existing office technology systems. The aim of this research is to use user-centered design to develop and evaluate an online multidimensional assessment tool for young people and GPs in the general practice setting.

Categories and Subject Descriptors

H.5.2 [User Interfaces]: User-centered design; J.3 [Life and Medical Sciences]: Health

Keywords

Screening, adolescent, user-centred design, general practice, therapeutic alliance

1. OVERVIEW

Adolescence is an important developmental period for physical, cognitive, social and emotional development. During this time, a wide range of health problems and risky health-related behaviours often emerge and co-occur, and persist into adulthood. These include high rates of substance use, mental illness, unsafe sex, and inadequate diet and exercise. Despite the high prevalence and co-occurrence of mental health problems and risk-taking behaviours, most young people do not seek help. However, 70-90% of young people regularly attend general practice, and whilst these consultations most commonly occur for physiological reason, this regular attendance suggests that General Practitioners (GPs) are in the unique position to deliver health prevention and promotion as part of young people’s routine health care, providing early intervention and referrals where necessary. A quick and effective way for GPs to identify areas of concern for young people is through a self-administered multidimensional screening or assessment tool. Completed prior to a consultation, this has the advantage of leaving more time to discuss relevant issues. Further, young people and GPs alike report that screening tools provide an opportunity to discuss sensitive topics that would not normally be brought up in the typical consultation [1]. Electronic screening tools, completed via computer or tablet, are particularly accepted by young people, with research suggesting they are more likely to disclose online compared with pen and paper [3]. They also have the benefit of functionality that sends summary reports and recommendations immediately to the GP. However, there is currently little research into how screening tools can be designed in a way that places young people as more active participants in their consultation and, in turn, how this might effect the therapeutic relationship. In addition, despite the advantages and need, the rate of screening for young people in general practice remains low [2]. There is a paucity of research that explores the needs of GPs, such as the form and content of screening reports and how the tools need to be designed in a way that enables them to be successfully integrated into a range of existing technology and office systems. The aim of this research is to use user-centered design to develop and evaluate an online multidimensional assessment tool for young people and GPs in the general practice setting. Specifically, the following research questions will be considered: What are the barriers and facilitators to screening adoption in general practice? How can technology be designed and implemented to overcome these adoption barriers and increase usage? How does completing an online multidimensional assessment tool effect the therapeutic alliance between the GP and young person?

2. REFERENCES


Help me keep my stuff safe

[Extended Abstract]
Sarah Ellen Webber
University of Melbourne
webberse@student.unimelb.edu.au

ABSTRACT
This paper reports on research into the design of an online repository for the personal possessions of young people in care, and resulting design heuristics for supporting collaboration around digital content in highly dynamic social contexts.

Categories and Subject Descriptors
H.5.m. [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Keywords
Boundary regulation, personal information management, identity

1. INTRODUCTION
For young people growing up in out-of-home care, many obstacles exist to the accumulation of personal artefacts as a resource for constructing and representing identity [2]. It is proposed that an online repository (a 'virtual locker') could provide continued access to information and possessions as young people transition through care placements and through relationships with multiple carers. This research explores how the virtual locker can be designed to support effective interaction and collaboration around personal possessions [3] in an environment characterised by impermanence and uncertainty in relationships.

2. RESEARCH DESIGN
Adopting a design science research methodology [5], this research was conducted in the context of a six-month project to investigate and develop a protoof-concept virtual locker. Initial focus groups with professional carers ('workers') and young care leavers were conducted to develop understandings of behaviours, motivations and tensions in managing the personal possessions of young people in care. Techniques were adopted from participatory design, including the use of characters, scenarios and conceptual sketches to enable participants to respond to prompts without discussing sensitive information or personal experiences.

Figure 1. Conceptual sketch of the Virtual Locker
Altman's theory of boundary regulation [1, 4] provided a framework for analysis of data gathered in these first investigations. Through this, key design objectives and corresponding recommendations were derived for the prototype virtual locker, which was constructed using digital wireframes and storyboards. A user evaluation was conducted with the focus group participants to understand the utility of the proposed prototype and specific design proposals. Participants were exposed to the prototype through demonstration and walkthrough, as well as hands-on task completion. This enabled capture of data through discussion, user observation and think-aloud protocol.

3. RESULTS
It was found that in this context critical design challenges lie in providing appropriate mechanisms for balancing a young person's control over their virtual locker with the need for carer oversight. Tools for collaboration and co-construction need to support this balance, and be flexible to accommodate the diverse and shifting circumstances of the young person. These objectives were translated into concrete design proposals for mechanisms which were incorporated into the prototype virtual locker.

The user evaluation indicated that several components of the prototype successfully address stakeholder concerns to provide a solution which would be effective for both young people and workers. This provides a basis for actionable recommendations for the design of the virtual locker. By generalising these findings we are able to propose design heuristics for supporting interaction around content in socially dynamic contexts. It is proposed that such systems should provide control for the individual, balance individual control with governance structures, and offer collaboration tools which accommodate shifting relationships.

4. CONCLUSION
Through investigating management of personal artefacts in out-of-home care, this research indicates that online collaborative systems can be designed to support users to interact around personal content despite problematic and highly dynamic interpersonal circumstances. Design heuristics to achieve this are proposed, drawing on the design and evaluation of the virtual locker. This suggests new opportunities to provide better support for collaboration around content in shifting contexts such as dynamic family structures, and for improved mechanisms for managing personal disclosure in online environments.

5. REFERENCES
Understanding the user experience of running with mixed reality stories
[Extended Abstract]
Alexander Kan
Department of Computing and Information Systems
The University of Melbourne, Victoria 3010, Australia
kana@student.unimelb.edu.au

ABSTRACT
The notion of running with mixed reality stories (MRS) has emerged in recent years. Although different aspects of these MRS have already received interest from researchers, to date still little is known about how MRS work as a whole in the context of running experience. This project aims to provide a holistic understanding of MRS and their effects on the experience of running.

Categories and Subject Descriptors
H.5.2 [Information Systems] User Interfaces

Keywords
User experience, mixed reality, narratives, game, running

1. BACKGROUND
If a story is “the representation of an event or a series of events” [1], a mixed reality story (MRS) is a story that creates a mixed reality by combining together virtual and physical environments. The most prominent example of this concept is a MRS called Zombies, Run! In this MRS a user is expected to actually run, while listening through headphones to a fictional story on zombie apocalypse. The main character in the story is the user herself. A runner experiencing Zombies, Run! runs not only in the physical world, yet also in a post-apocalyptic virtual world of the MRS. In that way running with Zombies, run! is different from running while listening to a “regular” audiobook.

Since its release in February 2012 Zombies, run! has gained more than 750 000 users, despite its price of $8 (the average price of an app in the AppStore is less than $3). It had also been the worldwide #1 Top Grossing Health & Fitness app in the AppStore. The success of Zombies, run! makes it reasonable to expect new MRS being created for running in the nearest future, as long as the evolvement of existing ones.

While some researchers already have explored certain aspects of running with MRS (e.g. [2]), to date still little is known about how MRS work as a whole in the context of running experience. This projects aims to provide a better understanding the user experience of running with mixed reality stories. To address the research questions of the projects we plan to use Benford and Giannachi’s model of user experience with mixed reality performances [3] and extend it, so it could include mixed reality stories as well.

Main research question: What is the user experience of running with mixed reality stories?
Sub-question #1: In which ways mixed reality stories support running?

Sub-question #2: How to create stories for mixed reality running apps?
Sub-question #3: How do actually runners perceive these stories?

2. RESEARCH DESIGN
2.1 Study 1: Explore
In the first study we explored how a story-based mixed reality app Zombies, run! can support running. We deployed the app to 11 runners, asked them to use it during the period of three weeks and then interviewed them about their experience. We found that the app changed running in three major ways. Firstly, it changed the way runs were organised. Secondly, it shook up established running routines. And lastly, it shaped the meanings associated with running. In this study we also observed how runners were dealing with ambiguities in the story. It turned out that as long as runners are willing to accept the story, ambiguity can be used to enrich the experience.

2.2 Study 2: Build
In this study we have looked into the process of creating stories for mixed reality running apps. To do that, we recruited a number of writers and ask them to make stories in this format. We found that writers were able to tackle the constraints and create a variety of different stories. We highlighted tactics that were common for all writers (e.g. use of ambiguity, use of curiosity and use of familiarity) and discussed in which ways use these tactics was different from similar tactics in “traditional” storywriting. We also noted that in this study ambiguity again appeared as a major finding, although we did not ask writers to use it specifically.

2.3 Study 3: Evaluate
In this study we will recruit runners again, ask them to run with the created stories and conduct interviews with each runner right after their runs. Firstly, we want to make sure that the stories we have created in the previous study still support running in the same three major dimensions. Secondly, we also want to explore in more detail the experience of ambiguity in the stories as it was the major theme in both previous studies.

3. REFERENCES
Non-verbal Interactions in Domestic Video Conferencing

[Extended Abstract]

Behnaz R Yeganeh
Microsoft Social NUI Centre
University of Melbourne
brostami@student.unimelb.edu.au

ABSTRACT
This research aims to understand how the design of video conferencing technologies should be improved to support non-verbal activities in domestic video conferencing. Recent research shows that people are increasingly using video conferencing tools to go ‘beyond talking heads’ to engage in a range of activities with their remote family members and friends. These activities include show and tell, watching video programs, cooking and eating, etc. The findings of initial studies emphasise the importance of these non-verbal interactions, however they demonstrate the inability of current video conferencing tools to fully support them.

Categories and Subject Descriptors
H5.m. [Information interfaces and presentation (e.g., HCI)]: Miscellaneous

Keywords
Video communication, symmetry, domestic, shared experiences, non-verbal interaction

1. BACKGROUND
Living apart from families and friends, either temporarily or permanently, is sometimes inevitable. However, people often aim to remain connected to their family members and friends when circumstances impose distance upon them [1]. While they use different combinations of technologies to have remote connections, video conferencing tools have been widely used to support this connection among geographically separated people [3,4]. Moreover, research on domestic use of video conferencing tools has revealed its significance role on reinforcing family identity and promoting family values [1], as well as strengthening relationship and shared identity of participants [2]. Therefore, video conferencing tools play an important role in maintaining distance relationships.

Recent literature has revealed that people not only use video chat tools to converse with others, but they also use it to mediate other everyday activities [2,3,4]. These activities include sharing the experiences of cooking and eating, watching video programs, participating in social gatherings and running house tours. In this research, these are called non-verbal activities. Nevertheless, current technologies do not support these activities well [2]. In other words, special functional requirements regarding mobility, framing, zoomability, audio and activity versus ‘talking head’ view as well as non-functional requirements such as support for different activities with different goals (e.g. performing parallel activities, watching remote partner doing an activity, etc.), multiple stakeholders, special contents (e.g. newly bought objects, food, etc.), specific context such as place and time and particular way of doing are the reasons that existing video conferencing tools are unsuitable for engaging in shared non-verbal activities [3,4]. So, they have to be reconsidered to support newly emerged needs. There exists little work on understanding what types of non-verbal activities people do while video conferencing, how they do those interactions, and how video conferencing tools can better support them.

2. METHOD
The research question of this study is ‘How should design of video conferencing technologies be improved to support non-verbal activities in domestic video conferencing?’ Three studies have been designed to address the research question. Firstly, an exploratory study investigates current experiences of non-verbal interactions in domestic context. Then, the second study builds upon the results of first study and designs a new technology to improve and enhance video-mediated experiences. Finally, the third study examines the user experience of the developed technology.

3. ACKNOWLEDGMENTS
I would like to thank my supervisors A/Prof Frank Vetere and Dr. Bernd Ploderer for their support and guidance. This study is supported by the Institute for a Broadband-Enabled Society.

4. REFERENCES


Won’t you be my neighbour? Clustered borrowing in libraries as evidence for shelf browsing

[Extended Abstract]

Dana McKay
Interaction Design Group, Department of Computing and Information Systems
University of Melbourne
dmckay@swin.edu.au

ABSTRACT
In this paper, I describe a study designed to evaluate the impact of borrowing on borrowing patterns in physical libraries, with a view to understanding the importance of borrowing in online systems.

Categories and Subject Descriptors
H.5.m [Information interfaces and presentation (e.g. HCI)]: miscellaneous

Keywords
Browsing, libraries, information seeking, log analysis books

1. INTRODUCTION
Browsing is a fundamental part of the human information seeking process [8], yet little work has investigated how best to support browsing online. Consider the case of books: little is known about the selection process [10], yet it is clear from both user report [3, 7] and early research [2] that browsing is important. Given the increasing uptake of ebooks, it is important to understand what readers need to support them in their book decision making online. This presentation will define browsing as it relates to books, give a brief overview of the literature in the field, and describe a recent study intended to establish the prevalence of browsing in academic libraries.

2. WHAT IS BROWSING?
Bates defined browsing in 2007 [1] as the visual investigation of a large scene, followed by sequential close examination of objects of interest within that scene with the continued visual presence of the overall scene. By that definition, library shelves are almost ideally designed for browsing—indeed recent work has sought to replicate them exactly [4]. Users report valuing the library shelves [7], often so much that their utility is a deterrent to ebook use [3].

Use of the shelves for browsing disrupts the foremost model of document triage [5] where the number of documents under consideration reduced at each stage: shelf browsing dramatically increases the number of documents under consideration after search. Even so, 55% of library users in a recent study report actively using the shelves, in part because this approach supports serendipitous discovery [7]. In contrast to the anecdotal evidence, the statistical evidence for shelf browsing is limited to a pair of studies from 1993 [2, 6], each of a relatively small number of users. The study identified here [9] aims to address that gap.

3. QUANTIFYING BROWSING
This study uses data from a large publicly available dataset to determine whether there is spatio-temporal clustering of book loans. In six sample US academic libraries, a clear neighbour effect is demonstrated—books near each other are likely to be borrowed on the same day. While, given the limitations of the data, it is not possible to establish with 100% certainty that this neighbour effect is specifically a result of the disruptive browsing described above, the literature on the use of library shelves [2, 4, 7] suggests browsing is a very likely reason for it. The importance of browsing in physical spaces suggests a need for better support for browsing online; how best to provide that support cannot be identified from the type of data driven study presented here and remains a task for future work.

4. ACKNOWLEDGEMENT
This paper contains information from OhioLINK Circulation Data (http://www.oclc.org/research/activities/ohiolink/circulation.htm) which is made available by OCLC Online Computer Library Center, Inc. and OhioLINK under the ODC Attribution License (http://www.oclc.org/research/activities/ohiolink/odcby.htm).

5. REFERENCES
Abstract
This research explores the use of tangible interfaces to support and enrich tele-consultation between a patient and health professional.

Categories and Subject Descriptors
H.5.2 [Information Interfaces and Presentation]: UI

Keywords
Tele-health, tele-consultation, tangible interfaces

1. INTRODUCTION
Tele-health is an umbrella term that encompasses a range of activities such as consultation, assessment, intervention, and health maintenance between a patient and the health supplier (such as practitioner) over a distance [4]. Tele-health serves as a vehicle to deliver easy and convenient access to health services to patients who otherwise have no other way to avail them such as people living in rural/remote areas and people with disabilities; which in turn makes the recovery and progress tracking easier for patients. Additionally, tele-health services provide financial benefits in terms of saving office hours, transportation cost and other miscellaneous costs of food and accommodation while visiting a health professional. Current tele-health services generally utilize video based communication to support tele-consultation sessions between the patient and health professional. However, such a setup demands technical expertise from both patient and provider, which they may not necessarily have; in turn would require consistent help and technical support. As a result, such systems have failed to go beyond the pilot phases [5]. Previous works [2,5] suggest that technology acceptance between the patient and health professional is a major factor behind the low commercial adaptation of the telehealth services. Moreover, while telehealth is related to human communication, yet human aspects of communication are greatly overlooked by the researchers [1]. For instance, it is essential to find out satisfaction and confidence of a patient in conveying his issues to the health professional while using the given telehealth system. For a successful tele-assessment, a doctor should get complete information and details of patient’s problems, which is possible only if the patient is equipped with the right technology as argued by Jang [3]. Motivated by these factors, I find an opportunity to support and enrich a tele-consultation session between a patient and health professional.

2. APPROACH
Major goal of this research is to explore the richness of tangible interfaces to capture and convey physical interactions between a patient and health professional during tele-consultation. My aim is to understand how physical interactions can be captured, shaped and conveyed to the other end in order to accurately mimic the traditional health care setting. For example, to support physical interactions of Alex, a patient for his physical assessment, a physical avatar (3D miniature replica) can be given to both the patient and health professional as shown in Figure 1. Instead of touching and exposing his own body parts, Alex can now touch the avatar to effectively communicate which parts of his body is suffering. His physical actions are mapped to the physical avatar at both the ends: for example, based on the touches, physical avatar will glow on specific touch locations. My hypothesis is that such an artificial setting yet enabling natural interactions can be one possible way to communicate physical interactions between a patient and health professional in a tele-consultation session.

Figure 1: Patient and health professional tele-consult via a shared 3D replica of the patient.

To support my research, firstly, I will understand the ways in which real and tele-consultations happen between patient and health professional through observational studies; which will provide me insights of the varied physical interactions required to communicate in a tele-consultation session. Next I will design and deploy a research prototype in various tele-consultation sessions to investigate its use. These observations will be supported through semi-structured interviews with participants (patients as well as health professionals) at the end.

3. REFERENCES
Experiences with Mobile Mental Health Therapies

[Extended Abstract]

Fernando Estrada
University of Melbourne
festrada@student.unimelb.edu.au

ABSTRACT
Mental health disorders have been identified by the World Health Organisation (WHO) as one of the leading causes of disability worldwide. The WHO has raised concerns about how those suffering with mental health disorders (users) may be assisted. As a resource tool, mental health therapies (MHTs) using mobile phones (mobiles) have become increasingly popular to reach users. However, despite their rapidly increasing numbers, high levels of attrition and/or unacceptability have been reported. In this regard, the aim of this research is firstly to understand the user experience of mobile MHTs. Secondly, develop a theoretical evaluation and testing model of mobile MHTs, and thirdly, to present practical recommendations for the design of mobile MHTs based on engagement and user experience.

Categories and Subject descriptors
H.4.0 [Information Interfaces and Presentation]: General

Keywords
Mental Health, mobile, user experience, acceptability, model

1. INTRODUCTION
The WHO identifies mental health disorders as one of the leading causes of disability worldwide [1]. Furthermore, Blanchard [2] states that at any given time 25% of the world population may be experiencing a mental health disorder. This is at a time where resources for face to face consultations are diminishing [3]. Consequently, alternative resources, such as those available by mobile, are crucial if an individual is to be supported to achieve a state of well-being.

Recent research on the use of mobile phones [4] suggests that 75% of people worldwide own and are active users of mobile phones. This also applies to individuals with mental health disorders (users) as claimed by Palmier-Claus [5]. As a result of this penetration of the market, and the affordability and ease of use mobiles may well present an efficient tool to support users.

2. MOBILES IN MENTAL HEALTH
A wide range of commercial mobile MHTs and specially tailored mobile MHTs for use within private clinical practice have emerged. However, despite their ever-increasing number, several concerns have been raised about the use of, and engagement with, mobile MHTs [6]. Commercial mobile MHTs may provide little or no research evidence about engagement [7] [8]. Secondly, research-based on MHTs using technology with random control trials (RCT) cite high ‘drop-out’ numbers [9]. Thirdly, clinical evaluation and testing has mainly considered the usability of such applications [8], therefore leaving a gap in relation to the user experience of mobile MHTs [5].

3. USER EXPERIENCE
Wright [10] suggests technology needs to go beyond cognitive approaches and usability tests, to address the user’s senses, intellect and feelings as part of an ‘aesthetic user experience’. Additionally, Battarbee [11] suggests that experience may be ‘co-experienced’, that is, when feedback from others influence the user’s interaction with technology.

4. RESEARCH DESIGN
Mobile MHTs researchers suggest participatory design [12] [13] and ethnography[14] [15] as suitable approaches to conduct research. However, due to the sensitivity of mental health [3], these approaches may need to be adapted. In this regard, the aim of this research is not only to explore the user experience of mobile MHTs but to create a model to evaluate and test them. This research will consider four studies. Each study will provide in-depth information for the following study. The first study will look at the experience of mental health professionals with mobile MHTs. The second study will consider the users’ experiences as presented on social media, blogs and forums. The third study will consider users’ experiences with mobile MHTs as well as organisations responsible for addressing protocols for mobile MHTs. The fourth study will examine an in-depth experience of users. Finally, this research will bring together these findings in order to generate a model to evaluate and test mobile MHTs.

The approach of this research is to further Wright’s ‘aesthetic user experience’. To do so, in the first study I will conduct semi-structured interviews with mental health professionals in order to understand their experience and ‘co-experience’ of mobile MHTs. In the second study I will obtain an understanding of users’ experiences of mobile MHTs as reported on social media platforms through the use of social media search engines and cognitive computing software. The experiences in both studies will be analysed to identify themes and categories as well as to define participants and any limitations that need to be considered on the subsequent studies. The third will consist of semi-structured interviews with users and organizations responsible for addressing guidelines for mobile MHTs. The forth study will consider a structured an in-depth interviews with users of mobile MHTs.

5. CONCLUSION
An agenda continuing the path of Wright’s ‘aesthetic experience’ [10] is needed if we are to understand the acceptability of mobile MHTs. Furthermore, an in-depth understanding of the user experience of mobile MHTs through the users, mental health professionals and institutions related to mobile MHTs is required.

6. REFERENCES
Full references upon request.
Communication Technologies during Family Mealtime

[Extended Abstract]

Hasan Shahid Ferdous,
Dept. of CIS, University of Melbourne, Australia
hasan.ferdous@unimelb.edu.au

ABSTRACT
The aim of this research is to understand the use of communication technologies during family mealtime, whether they are used as information source, entertainment, or enabling social connection, among many other uses. We aim to understand the comprehensive relationship between technology and people, between the people themselves, and also the interaction among multiple technologies or devices in the context of family mealtime activities.

Categories and Subject Descriptors
H.5.m [Information Interfaces and Presentation]: Miscellaneous

Keywords
Family, Mealtime, Technology.

1. INTRODUCTION
The use of different digital communication technologies (e.g., cellular phones, computers, etc.) is very common in the context of a family nowadays. Yet family mealtime creates some unique opportunities regarding the use of communication technologies, as it is one of the few occasions that often recommend the presence of all the family members together at the same place and time. Some of the technologies are related to food preparation and consumption, while some others emphasize the social experience of commensality.

Television is probably one the very common technologies found near the dining place and a lot of researches, especially from sociological and health perspectives have discussed the role it plays in association with our mealtime activities. Other digital devices, especially cellular phones are widely used nowadays, and we know very little about if they have made any change in the way we enjoy our family meals. In this work, we want to identify the most commonly used communication technologies associated with our mealtime activities, where and how these devices are placed, used, and shared as an information source, entertainment, or enabling social connection, among many other uses.

2. RELATED WORKS
Grimes et al. [1] discussed the aesthetic aspects of mealtime, where they emphasized on the creativity in cooking, pleasure and nostalgia in the togetherness of family meals, relaxation, and food as a gift towards others. They also discussed the challenges of introducing technology to augment this ‘celebration’. Hupfeld et al. [2] took an ecological approach to uncover the role tabletops, spaces, and artifacts play in the social organization of domestic eating practices. They presented the challenges we need to consider in designing for home regarding independence and socialization, isolation and conviviality, reminiscence and identity.

One notable approach is ‘4Photos’ by O’Hara et al. [3], where they emphasized the use of situated photo mementoes within the social concerns of culturally specific rhythms, norms, rights, and responsibilities. They showed that these personal memory resources could be invoked within the context of social work being done at a shared meal.

3. INITIAL FINDINGS AND CONCLUSION
We are at the beginning stage of our first study and got our ethics application approved. Using semi-structured interviews, we aim to understand the current practices of using different sorts of communication technology around our family mealtime. Then we will try to capture the family mealtime activities in real-life settings using video recordings. We will analyze the recordings based on grounded theory approach, in order to find the main themes.

From our discussions in informal user forums about the use of technology during mealtime, we have revealed several common themes and would like to focus on them in our first study. First of all, there are clearly two different stands about using technology during mealtime. In some families it is strictly prohibited, while others maintain an open position regarding this. It has strong familial influence and the family culture is almost always followed in using (or not using) technology during mealtime. Television is the most commonly used device, while mobile devices are slowly gaining access to our dining space. Many of the families keep their TV on during mealtimes, while not watching any particular program or paying any attention to it. How this device went from foreground to background, is of special interest to us.

Understanding the family practice in using and sharing the communication devices during mealtime can enable us to design novel interaction techniques to support the experience of commensality in our day to day family life. Supporting the diverse family practices and scenarios in this regard remains a major research challenge.

4. REFERENCES
Screen Ecologies and the Future of Domestic Gaming

[Extended Abstract]

Marcus Carter  
Interaction Design Lab  
The University of Melbourne  
marcusc@unimelb.edu.au

Bjorn Nansen  
Interaction Design Lab  
The University of Melbourne  
nansenb@unimelb.edu.au

Martin Gibbs  
Interaction Design Lab  
The University of Melbourne  
martin.gibbs@unimelb.edu.au

ABSTRACT
In contrast to conceptualizations of gaming as immersive, fully-engaging experiences, contemporary domestic play occurs across a variety of fluctuating frames of activity and registers of engagement. In this talk, I will show how this is attributable to what I call the domestic screen ecology; the emerging coexistence and simultaneous engagement with activities on multiple monitors, televisions, laptop computers and smart phones. I argue this has considerable consequence for the future design of engaging interactive experiences and challenges the fallacy of immersion purported by new VR platforms like the "Oculus Rift.

Categories and Subject Descriptors
K.8.0 [Personal Computing]: General - Games

Keywords
Screen ecologies, domestic gaming, multi-gaming

1. INTRODUCTION
As computing technologies have grown increasingly ubiquitous, a proliferation of screens has occurred. In addition to the widespread preference amongst gamers (and non-gamers) for multiple monitors; televisions, laptop computers and smart phones have all invaded the domestic personal computing space. Rather than a single screen sitting on a table capturing the full attention of the user, the modern computing experience is commonly much more complex and distributed across multiple devices. The user's engagement with any or all of these screens varies depending upon context or activity. This is the contemporary screen ecology.

Figure 1. An example of a contemporary screen ecology

2. RELATED WORK
The use of multiple screens during computer use has been long-noted in HCI, both in the context of professional work and non-gaming leisure, such as television viewing. However, Brown et al. [1] note that "there has been very little research systematically investigating how attention is split during dual-screen human-computer interaction" [p. 666]. Studies which do exist have only studied interrelated applications; for example companion applications for TV viewing, or single games that utilize two screens at once. No research exists that has investigated the use of multiple screens for distinct, simultaneous leisure activities, and how they interrelate.

3. MULTI-GAMING
As a result of this complex media environment, our research is revealing how digital games are often consumed in parallel with other media such as TV, YouTube, social sites like reddit or Facebook, serious work and student learning. However the most interesting emergent practice is multi-gaming: the simultaneous play of distinct digital games. For example, playing passive games like Sim City in between deaths in Call of Duty. Our research so far [see 2] has identified genres of engagement modalities in digital games that successfully interplay for multi-gaming, suggesting a rich site for future design.

4. LESSONS FOR ENGAGEMENT
Brendan Keogh has argued against the "fallacy of immersion" in games studies which privileges the content of games over the context in which it is experienced [3]. Rather, the actual, material, cultural and formal properties of games are significant for understanding, critiquing and designing them. By examining the contemporary screen ecology and the phenomenon of multi-gaming, this research is demonstrating how ‘engagement’ can still be meaningful beyond measures of immersion and quantity of use, and how this is relevant to successful game design.

5. THE FUTURE OF DOMESTIC GAMING
Packaging on the (commercial failure) 1994 VFX1 virtual reality headset purports to keep “the real world out and the virtual world in”. By demonstrating how concepts like ‘immersion’ obscure our understanding of how players engage with and enjoy digital games in the context of the real world, our research suggests one possible explanation for why VR platforms have broadly failed. Rather than immersive, ‘matrix’ style gaming purported by these devices, we envision a digital gaming future that is played across screens, devices and contexts in addition to other activities. The goal should not be to keep play apart from everyday life (in the virtual, not the real) but embedded within.

6. REFERENCES
Software Engineering

Loop Untangling

[Extended Abstract]

Kathryn Francis
The University of Melbourne CIS Department, and NICTA Victorian Research Laboratory
k.francis@student.unimelb.edu.au

ABSTRACT
This paper concerns the translation of program code into constraints (mathematical equations defining the effects of the code), in order to facilitate automated reasoning about program behaviour. In particular a new technique is presented for the translation of bounded for and while loops.

Categories and Subject Descriptors
F.3.1 [Logics and meanings of programs]: Specifying and Verifying and Reasoning about Programs

Keywords
Loop unwinding, constraint model

1. INTRODUCTION
A translation from code into an equivalent constraint model allows a constraint solver to use sophisticated techniques to efficiently search over inputs to the program looking for desirable or undesirable behaviour. There are many applications for such automated reasoning. For example, it is beneficial to be able to prove that critically important software (e.g. that controlling rocket engines or life support systems) does not contain certain types of errors.

When performing the translation from code into constraints, loops (such as for loops and while loops) are a significant problem. Program variables will take different values in different iterations of the same loop, so it is not possible to use a single mathematical variable to represent the value on a given line of code. For bounded loops (those where a limit on the number of iterations is known or can be assumed), the usual technique (used in e.g. [1, 2]) is to create a copy of the loop body for each potential iteration. This is called loop unwinding.

The problem with loop unwinding is that it unnecessarily ties the execution order of iterations to their identities. Copies of the loop body are identified as representing the iteration performed first, second, third, etc. This often results in complicated constraints for each individual iteration, as the values of key expressions are unknown.

This work investigates an alternative, called loop untangling, where copies of the loop body are instead identified by the value of a key expression. For example, if the chosen expression $e$ is known to take value $A$ or $B$ in the first iteration, and value $B$ or $C$ in the second iteration, then we will create 4 copies of the loop body, each having a known value for $e$. Both $A$ and $C$ can only occur once, so we create a single copy for these. $B$ may occur twice, so we need two copies - one for the first time $e$ takes value $B$ and another for the second time this potentially occurs. This approach results in (sometimes many) more copies of the loop body, but because the value of $e$ is known, the constraints for each copy can be greatly simplified.

To produce a correct constraint model it is still necessary to link the iterations together appropriately so that they represent a valid execution path through the loop. To achieve this we extend the translation technique described in [3] to include an explicit representation of the execution path.

2. PRELIMINARY RESULTS
Experiments using two parameterized examples from [3] have shown significant improvements over standard loop unwinding. The table below gives the solver execution time and number of dead ends encountered when solving models produced using loop untangling versus standard loop unwinding. Each figure is the average for 30 random instances of the given size, with the number of instances reaching the 10 minute timeout shown in brackets.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Solving Time (secs)</th>
<th>Dead Ends (000s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unwind</td>
<td>untangle</td>
</tr>
<tr>
<td>pizza</td>
<td>4</td>
<td>94.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>320.7 (13)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>470.1 (22)</td>
</tr>
<tr>
<td>routing</td>
<td>5</td>
<td>12.9</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>102.8</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>569.3 (20)</td>
</tr>
</tbody>
</table>

3. ACKNOWLEDGMENTS
NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. Kathryn Francis is supervised by Peter Stuckey and Harald Søndergaard.

4. REFERENCES
Automated debugging using program spectra

[Extended Abstract]

Neelofar
University of Melbourne
neelofar.eme@gmail.com

Lee Naish
University of Melbourne
lee@unimelb.edu.au

Rao Kotagiri
University of Melbourne
kotagiri@unimelb.edu.au

ABSTRACT
This paper presents an automated approach for software fault localization using program spectra obtained by dynamic execution of a software system. We present few metrics obtained by genetic programming to rank program statements or blocks according to how likely they are to be buggy. The metrics are tested on 92 faults from four unix utilities and the results shows that they outperforms other metrics, such as Jaccard, Trantula and Ochiai in single bug program.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; D.2.5 [Software Engineering]: Testing and Debugging

Keywords
Automated debugging, program spectra, fault localization

1. INTRODUCTION
Debugging software is very important but resource intensive tasks in software engineering [4]. Fault localization is the most expensive among the activities of debugging. Any improvement in this domain will automatically improve overall debugging process [3].

Spectra-Based fault localization technique uses execution traces of software program (program spectra) to find the likelihood of each statement being faulty.

The performance of a fault localization technique depends mainly on risk evaluation metrics [1]. Many risk evaluation formulas or suspiciousness metrics are defined in literature that ranks the program statements on the basis of their execution profile. However all of these risk evaluation metrics are designed by human intuition or inherited from other fields. In this research we are evolving risk evaluation formula from program spectra directly using genetic programming.

2. EXPERIMENTAL RESULTS
Detail of the operators used in our experiments is given in table 1. ep, np and ef are executed passed, not executed failed and executed failed values from program spectra.

Table 1: Genetic Program Operators

<table>
<thead>
<tr>
<th>Operator</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiply</td>
<td>a*b</td>
</tr>
<tr>
<td>Divide</td>
<td>a/b if b&gt;0; a/.001 if b==0</td>
</tr>
<tr>
<td>Add</td>
<td>a+b</td>
</tr>
<tr>
<td>Subtract</td>
<td>a-b</td>
</tr>
<tr>
<td>Opt</td>
<td>Optimal metric [2] ef + ϵ np ; ef - ϵ ep</td>
</tr>
<tr>
<td>W73</td>
<td>0.7 ef + 0.3 np</td>
</tr>
</tbody>
</table>

Table 2: GP Generated Formulas

<table>
<thead>
<tr>
<th>Formula</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>((ef Opt ef) Opt (ef Opt ef)) Opt (ef Opt np)</td>
<td>11.6445</td>
</tr>
<tr>
<td>((ef Opt ep) Opt ((ef Opt ef) Opt (ep Opt ef)) Opt (ef Opt ep))</td>
<td>11.1665</td>
</tr>
<tr>
<td>(ef w73 ep) w73 ((((ep w73 ef) w73 ef) w73 ep) w73 ef)</td>
<td>12.4076</td>
</tr>
</tbody>
</table>

is 11.6445 and 12.4076 which proves that derived formulas are performing better as compared to Trantula and Jaccard. Same tie breaking scheme is applied to the ordered list of statements after applying Jaccard, Trantula and derived formulas.

3. REFERENCES
Review of Interdependence in Coactive Design

[Extended Abstract]

Ronal Singh
Department of Computing and Information Systems
The University of Melbourne, Victoria, 3010, Australia
ronals@student.unimelb.edu.au

ABSTRACT
Interdependence is defined as the set of complementary relationships between two or more parties. This paper reviews the conditions that lead to interdependence between agents and highlights that there is a need to further investigate it.

Categories and Subject Descriptors
H.1.2 [Models and Principles]: User/machine systems

Keywords
Interdependence, levels of automation, coactive design

1. INTRODUCTION

Levels of automation (LoA) were introduced to describe how machines and humans are assigned control of a system and Miller [3] has recently reviewed a number of LoA models. Johnson et al. [2] highlighted that these models do not adequately capture human machine teamwork, because machines should be able to work unsupervised when needed as well as be capable of joint activity with human operators. Furthermore, not only should humans and machines be able to hand-off tasks to each other but they should also be able to participate on tasks requiring collaboration [2].

To design systems capable of exhibiting human-agent collaboration, Johnson et al. [1] proposed the Coactive Design Method, which allows designers to identify, analyse, implement and evaluate interdependence relationships between agents. The term “coactive” is used to emphasise that humans and agents will work closely and require continuous interaction. They highlight interdependence as the central organising principle of Coactive Design Method.

This paper describes interdependence as a concept and emphasises the need to further investigate interdependence.

2. INTERDEPENDENCE

Interdependence, in the context of joint activity, has been defined as “the set of complementary relationships that two or more parties rely on to manage required (hard) or opportunistic (soft) dependencies in joint activity” [1].

Understanding an interdependent relationship requires understanding the reason for the dependencies between agents and mechanisms of supporting it. Agents can be dependent on each other because of the lack of capacity or reliability, that is, because agents cannot complete tasks independently. The agents should have coordination mechanisms to support the interdependent relationship, that is, the agents should have the ability to ask for and receive assistance from other agents. The relationship is complementary because agents are dependent on each other. The context matters, that is, in addition to each agent having its own situation, there is also a need to be aware of the situation enveloping other parties. So the interdependent relationships between parties determine what is relevant, and thus help identify what common ground [4] should comprise.

3. DISCUSSION

Johnson et al. [1] define interdependence in the context of joint activity. The study of examples where joint activity is not necessary for interdependence to exist between agents presents an important question:

What conditions lead to interdependence between agents?

To answer the above question, there is a need to further investigate interdependence. The aim of this study is to have a better understanding of interdependence, which may lead to a better understanding of the mechanisms of supporting interdependence in the context of human agent collaboration. One of the outcomes of this research is our semi-formal definition of interdependence [5].

4. ACKNOWLEDGMENTS

I sincerely thank my supervisors, Prof Liz Sonenberg and Dr Tim Miller for their valuable feedback and guidance.

References
Download the program and proceedings from our website!


cis-dc@unimelb.edu.au
Analysis of Road Traffic Using Contrast Mining and GPS Trajectories

Xiaoting Wang, Christopher Leckie, Tharshan Vaithianathan

Introduction
Traffic congestion has been one of the most critical issues in many modern cities. A particular problem of interest is to characterize the traffic flow before and after an event, such as road closure, and assess the impact of this event. This study uses contrast mining techniques to identify emerging patterns in road traffic that are associated with an event. The GPS trajectories of more than 10000 taxis in the city of Beijing was used [1]. By computing the "Growth Rate" of traffic on each road segment, the road closure event can be detected. We have developed methods to extract a graph from road networks and techniques to process and mine taxi GPS trajectories.


Road Network Extraction
The road network surrounding a road closure event in Beijing (Figure 1) was downloaded from OpenStreetMap.org. Simplification of the road network was performed using QGIS (www.qgis.org) and GRASS GIS (grass.osgeo.org) to extract main roads and main intersections. The road network was converted into a node-edge list, where nodes = intersections and edges = roads.

Contrast Mining Using GPS Trajectories
Contrast Mining was initially used in analyzing supermarket transactions [2]. Taxi trajectory data from more than 10000 taxis were grouped based on a road closure event. Dataset One (D1) contains trajectories in days \{1,2,3,4,5\}, which are the days before the road closure. Dataset Two (D2) contains days \{6,7\}, which is the days when a major road in Beijing was closed due to the Chinese New Year. Drawing 4 hours of data in one day from D1 and another 4 hours at the same time of the day from D2, we compute the Growth Rate (GR) of traffic on each edge (road):

\[
\text{Growth Rate of Edge } A \text{ (day1 to day2)} = \frac{% \text{ of Taxis Traveled Through Edge } A \text{ in Day 2}}{% \text{ of Taxis Traveled Through Edge } A \text{ in Day 1}}
\]

The mean of GR is nearly constant at 1 and std = 0.4 for all days (Figure 2, 3, and 4). Outside this range, we can identify roads with growing traffic (Blue) when:

GR > mean + std

and roads with reducing traffic (Green) when:

GR < mean - std

As a result, the road closure event was successfully detected.


Conclusion and Future Work
We successfully detected the road closure event in Beijing using the GPS trajectories of more than 10000 taxis and contrast mining techniques. Future work will be to assess the impact of the road closure on its nearby roads and determine the changes in the choice of driving directions due to the closure.
Answering Complex Questions in Neuroimaging Informatics

Aref Eshghi Shargh\textsuperscript{1}, Simon Milton\textsuperscript{2}, Andrew Lonie\textsuperscript{3}, Gary Egan\textsuperscript{4}

\textsuperscript{1} CS Department, The University of Melbourne
\textsuperscript{2} School of Medicine, Monash University
cjfe@students.unimelb.edu.au, Tel: +61 3 8341 6648

Introduction and Background

Neuroscience is a broad and important area of study and research, motivated by both a desire for basic understanding and implications for health. Neurological diseases comprise one of the largest groups of the human diseases with more than 2400 individual disorders\textsuperscript{1}. Large neuroimaging\textsuperscript{2} (as large as 2 GB) are to be studied in neuroscience research. Our focus is on answering complex questions posed on annotated neuroimages. Annotations record metadata about an entity\textsuperscript{3}.

We are using ontologies as a tool. Ontologies are formal, explicit specification of a conceptualization of a field\textsuperscript{4} and they can be seen as maps of different fields.

Application

As an example, we propose and solve a question: ‘Calculate the volume of sub parts of the white matter of inferior parietal lobe of the subject’. Based on the semantic circle and pyramid, this is a Level 4 question:

There are many challenges involved in this simple example. The annotation file does not show the parts that we are looking for. This is due to neuroimaging visibility issues (they do not show parts smaller than a certain size). So, we use the FMA ontology to find these sub parts, using the following code:

Still, we are not able to find the sub-part’s volumes, because the annotations follow a different terminology from the FMA ontology. To resolve this issue, we use another ontology which is called Foundational Model of Neuroanatomy (FMN) to map the annotations to the parts in the FMA.

Having the sub-parts (code A) and being able to find them in the annotation file (code B), we can retrieve the volume of each part and calculate the overall volume.

Conclusion

The overall process of answering a sample complex level 4 question is as follows:

Audience Experience in Domestic Videogaming

Videogames are often played socially, but not everybody in the room actively plays. This thesis explores the experience of audience members who are watching others play.

**Study 1**

**RQ:** What social interaction takes place around physical console gaming?

- **8 families**
- **2 weeks**
- **Recorded gameplay**

**Durable roles**

- Bystander ➔ Audience ➔ Player

**Ephemeral audience roles**

- Spectator
- Heckler
- Commentator
- Director
- Advisor
- Orchestrator
- Journalist
- Shadow player
- Supporter
- Puppeteer
- Cheerleader
- Coach

**Study 2**

**150 participants (pairs)**

**Lab experiment**

**Aim:** Explore effects of turn anticipation and game physicality level on audience experience.

![Graph showing enjoyment ratings](image)

Conclusions:

- Playing more fun than watching
- Anticipatory play effect
- Residual play effect
- Kinect games most fun to watch

**Study 3**

**24 participants**

**Lab experiment**

**Aim:** Explore how attention, game physicality level and social context interact to influence audience experience.

**Method:**

- Split screen videos
- Eye gaze tracking
- Attention/experience ratings

**Conclusions:**

- Attention a poor predictor of experience
- Weak correlation of subjective and objective attention
- Physicality needs social context for enjoyment effects to emerge
- Regular social gamers found the situation more enjoyable

John Downs

PhD 2014

jpdowns@unimelb.edu.au
Data Quality and Data Cleansing in Bioinformatic Databases

Student: Qingyu Chen

Supervisors: Justin Zobel and Karin Verspoor

References:
- Peter G Korning et al, Cleaning the GenBank arabidopsis thaliana data set, 1996
- Christian Schonbach et al, Data warehousing in molecular biology, 2000
- Judice LY Koh et al, Duplicate detection in biological data using association rule mining, 2004
- Judice LY Koh et al, Correlation-based methods for biological data cleaning, 2007
- Taxiarchis et al, Secondary use of EHR: data quality issues and informatics opportunities, 2010
- Harris James. Can you bank on GenBank?, 2013

http://www.benjaminwicks.com/portfolio/

Bioinformatic Databases:
- vast and inconsistent
- GenBank, a DNA sequence database, contains 173,353,076 sequences in June, 2014 and doubles every 18 months
- Different error types make it even more complex: redundancy, strange sites, etc.
- Data quality is an ongoing problem in such databases

Redundancy Detection

The problem currently working on

Existing approaches:
- High similarity
- Duplication

Our approach: Clustering

Existing approaches:
- Association rule mining
- Data gathering
- Progress made:

Challenges:
- Different error types make it even more complex
- Redundancy
- Strange sites
- Data quality is an ongoing problem in such databases

Progress Made:
- Data gathering
- Pairwise alignment
- Redundancy detected from Entrez sample
- >50% redundant data detected from Entrez sample
- 15% erroneous data found in a GenBank dataset
- 17% erroneous data found in SALD
- >48% incompleteness found in clinical data sample
- >50% redundant data detected from Entrez sample

General database:
- DNA sequence database
- Entrez
- GenBank
- Incomplete
System Objective

Minimize the total carbon footprint of the Cloud provider through efficient VM placement.

Cloud broker makes the placement decision based on the:

- Data centers' power usage effectiveness (PUE)
- Energy sources' carbon footprint rate
- Servers' proportional power usage

Future Directions

- IT load and outside temperature
- Different user applications and VM holding times
- Inter-data centers network distance and data locality

Cloud Users: Send their VM requests to Cloud users.
EEG Cloud Broker: Receives users' requests.
EEG Cloud Information Service: Keeps detailed information of each cloud center.
EEG Cloud Information Service: Keeps energy and carbon-efficient (ECE) VM placement algorithms.

Cloud Provider: Consists of
- The cloud provider.
- The cloud broker.
- The cloud information service.

Cloud Provider: Consists of:

- Distributed data centers with different PUEs and carbon footprints.
- Application and data distribution.
- Energy and carbon-efficient (ECE) VM placement algorithms.

VM Placement Algorithm

VM placement can be seen as a bin-packing problem with different bin sizes (physical servers).

Proposed Energy and Carbon-Efficient (ECE) VM placement algorithm is based on a two-level decision-making model:

1. Selects data center/cluster with the minimum (carbon footprint x PUE).
2. Selects host with the minimum increase in power consumption.

VM Placement Policies:

- ECE
- First-Fit (FF)
- First-Fit Most-Full First (FF-MF)
- First-Fit Power-Efficient (FF-PE)
- Carbon-Efficient First-Fit (CE-FF)
- Energy and Carbon-Efficient (ECE) algorithms.

Performance Evaluation

- ECE and carbon-efficient (ECE) Cloud architecture.
- Results:
- Energy and carbon-efficient (ECE) Cloud architecture.
- Energy and carbon-efficient (ECE) Cloud architecture.
- Energy and carbon-efficient (ECE) Cloud architecture.
- Energy and carbon-efficient (ECE) Cloud architecture.
- Energy and carbon-efficient (ECE) Cloud architecture.

References:

Atefeh Khosravi and Rajkumar Buyya
Cloud Computing and Distributed Systems (CLOUDS) Laboratory,
The University of Melbourne
atefehk@student.unimelb.edu.au, rbuyya@unimelb.edu.au

Energy and Carbon-Efficient Resource Management for Geo-Distributed Cloud Data Centers

Melbourne School of Engineering

System Model Components

- Cloud Users
- Cloud Provider
- ECE Cloud Information Service
- ECE Cloud Broker

Introduction

Attefeh Khosravi and Rajkumar Buyya

Cloud Computing and Distributed Systems (CLOUDS) Laboratory,
The University of Melbourne

Melbourne School of Engineering
Fragment-based Planning using Column Generation

Toby Davies – PhD Student  Supervisors: A Prof Adrian Pearce; Prof. Peter Stuckey; A Prof Harald Søndergaard

The Bulk Freight Rail Scheduling Problem (BFRSP)

- A Service is a sequence of movements by a consist (a locomotive and wagons) over the rail network
- Objective: satisfy as many orders as possible with minimum sum of service durations

Key Technique & Advantages

- Linear programming, in particular the dual, allows us to accurately predict the cost of resource consumption—we use both the primal and the dual in feedback with planner
- MIP-Planner decomposition enables scaling to far larger problems, when compared to other anytime algorithms (see results)
- High-level temporal planning language is concise and easy to modify

Potential application domains: state-dependent resource constrained problems involving time (e.g. Transportation, Mine Scheduling, Production scheduling, etc.)

References


The Problem: motivated by industry experience

Resource constrained planning problems (e.g. BFRSP): known to be challenging to solve using current technology, even non-temporal ones

Industry experience: small modification to BFRSP problem took 9 months to amend the 10,000 lines of C++ code!

- High-level temporal planning languages: promises a more concise way of expressing the temporal domain, such as concurrent Golog De Giacomo et al. (2000); to solve larger problems
- Linear programming: for automated planning heuristics, Van den Briel et al. (2007) exploits primal solutions to LP

Challenge: Resource constrained temporal planning problems are typically beyond the scope of current planning technology!

The Solution: combines strengths of MIP & Planning

Automated translation into MIP/planning decomposition:

- Automated planner: Domain specific temporal planning language guides search for feasible action sequences (BFRSP services)
- MIP master problem: Global mixed integer linear program relaxation predicts costs (resource consumption) based on dual information

Potential application domains: state-dependent resource constrained problems involving time (e.g. Transportation, Mine Scheduling, Production scheduling, etc.)

Comparative Results: increasing network size & orders

<table>
<thead>
<tr>
<th>V</th>
<th>O</th>
<th>m/p</th>
<th>Golog</th>
<th>popf2</th>
<th>cpx</th>
<th>FBP</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>0.4</td>
<td>0.3</td>
<td>7.5</td>
<td>1.0</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>1</td>
<td>—</td>
<td>—</td>
<td>2.3</td>
<td>3.3</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>1.7</td>
<td>2.0</td>
</tr>
<tr>
<td>6</td>
<td>8</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>7.5</td>
<td>7.6</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>37.9</td>
<td>29.7</td>
</tr>
<tr>
<td>6</td>
<td>32</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>50.3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>64</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>150.3</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>128</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>418.9</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>256</td>
<td>2</td>
<td>—</td>
<td>—</td>
<td>588.7</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>44</td>
<td>11</td>
<td>—</td>
<td>—</td>
<td>315.0</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>88</td>
<td>11</td>
<td>—</td>
<td>—</td>
<td>363.0</td>
<td></td>
</tr>
</tbody>
</table>

BFRSP: Time to first solution for increasing track network vertices, $|V|$, and orders, $|O|$, in seconds (1800s time limit, 4GB memory)

FBP: Fragment-based planner Davies et al. (2014)

CPX: Constraint programming solver using Lazy Clause Generation Ohrimenko et al. (2009)

POPF2: A heuristic satisfying planner, Coles et al. (2010)

The object of this research project is to increase situation awareness (SA) in information security risk management (ISRM). Endsley’s SA theory holds that SA is developed in three progressive stages (perception, comprehension, and projection), and that SA occurs in the context of decision making leading to action on—or in response to—a particular situation. While SA theory has not previously been applied to organizations, a document-based case study of the U.S. National security intelligence enterprise (USNSIE) revealed that 1) SA theory can be used to explain the functions of the USNSIE, and 2) the USNSIE provides us a model of an organizational SA-support system, which can be adapted for organizations’ ISRM processes as illustrated below.

References:


1. The SET COVER problem

The SET COVER problem involves a universe $U$ of items $[0 \ldots n-1]$ and a family $F$ of sets $\{S_0, \ldots, S_{m-1}\}$, with $S_i \subseteq U$ and $\bigcup_{i=0}^{m-1} S_i = U$. The task is to find the smallest subset $A \subseteq F$ such that the union of all sets in $A$ covers $U$ as a whole [1]. In Figure 1, $\{S_0, S_1, S_2\}$ covers every item of $U$.

![Figure 1: An instance of the problem where $U = \{0, \ldots, 9\}$ and $F = \{S_0, \ldots, S_5\}$.]

The problem is applicable to a range of logistic problems. In one such layout scenario, the minimal number of new telephone poles (sets) can be computed from a large collection of potential pole locations to cover all homes (items) in a particular town.

2. Solving the SET COVER problem

The well-known greedy strategy [2] repeatedly adds sets with the biggest number of uncovered items to a growing solution, constructing a solution that is at most 1+$\ln n$ times as big as the minimal one [2].

Algorithm 1 The GREEDY approach

**Input:** Family $F$ of sets $S_i \subseteq U$, with $\bigcup_{i} S_i = U$

**Output:** Group $A$ of sets, with $\bigcup_{i \in A} S_i = U$.

1. $A \leftarrow \emptyset$, covered $\leftarrow \emptyset$
2. while $\text{covered} \neq U$ do
3. $S_i \leftarrow \text{argmax}_{S \in A} |S \setminus \text{covered}|$
4. $A \leftarrow A \cup \{S_i\}$
5. $\text{covered} \leftarrow \text{covered} \cup S_i$
6. end while
7. return $A$

Implemented as EAGER-GREEDY, the approach uses a priority queue to track of their priorities. Each priority is the number of uncovered items to a growing solution, constructing a solution.

Algorithm 2 The LAZY-GREEDY algorithm

**Input:** As for Algorithm 1

**Output:** As for Algorithm 1

1. $A \leftarrow \emptyset$, covered $\leftarrow \emptyset$, $\text{pqueue} \leftarrow \emptyset$
2. for $j \leftarrow 0$ to $m-1$ do
3. $\text{pqueue}.\text{insert}(|S_j|, j)$
4. end for
5. while $\text{covered} \neq U$ do
6. $(|S_i|, i) \leftarrow \text{pqueue}.\text{maximum}()$
7. $S_i \leftarrow S_i \setminus \text{covered}$
8. if $|S_i| = |S_j|$ then
9. $A \leftarrow A \cup \{S_i\}$ // all items in $S_i$ were clean
10. $\text{covered} \leftarrow \text{covered} \cup S_i$
11. else
12. $S_i \leftarrow S_i \setminus \text{some items in } S_i \text{ were covered}$
13. $\text{pqueue}.\text{insert}(|S_i|, i)$
14. end if
15. end while
16. return $A$

In Algorithm 2, the LAZY-GREEDY stores (insert) in the priority queue all $S_i$’s of initial sizes as their priorities, and iteratively retrieves (maximum) the putative largest $S_i$ from its queue. It then cleans $S_i$ and checks the size of $S_i'$. If the size is unchanged, $S_i$ is actually the largest, and is added to $A$. Otherwise, $S_i'$ is inserted back into its priority queue with its smaller size as its updated priority.

The EAGER-GREEDY algorithm ensures that all remaining sets have their correct priorities at all times, but LAZY-GREEDY updates the priority of a candidate set if and only if the set fails the size check.

3. Expensive on large data instances

There are at most $|F| = m$ additions to a solution, and at most $M$ priority updates, where $M$ is the total number of items in $F$. The overall time of the program on a data instance is dominated by the $O(M)$ cost of updating, assuming the time of each set update is $O(1)$ with the priority queue implemented as a fast array-plus-list.

Algorithm 2 The LAZY-GREEDY algorithm

**Input:** As for Algorithm 1

**Output:** As for Algorithm 1

1. $A \leftarrow \emptyset$, covered $\leftarrow \emptyset$, $\text{pqueue} \leftarrow \emptyset$
2. for $j \leftarrow 0$ to $m-1$ do
3. $\text{pqueue}.\text{insert}(|S_j|, j)$
4. end for
5. while $\text{covered} \neq U$ do
6. $(|S_i|, i) \leftarrow \text{pqueue}.\text{maximum}()$
7. $S_i \leftarrow S_i \setminus \text{covered}$
8. if $|S_i| = |S_j|$ then
9. $A \leftarrow A \cup \{S_i\}$ // all items in $S_i$ were clean
10. $\text{covered} \leftarrow \text{covered} \cup S_i$
11. else
12. $S_i \leftarrow S_i \setminus \text{some items in } S_i \text{ were covered}$
13. $\text{pqueue}.\text{insert}(|S_i|, i)$
14. end if
15. end while
16. return $A$

In Algorithm 2, the LAZY-GREEDY stores (insert) in the priority queue all $S_i$’s of initial sizes as their priorities, and iteratively retrieves (maximum) the putative largest $S_i$ from its queue. It then cleans $S_i$ and checks the size of $S_i'$. If the size is unchanged, $S_i$ is actually the largest, and is added to $A$. Otherwise, $S_i'$ is inserted back into its priority queue with its smaller size as its updated priority.

The EAGER-GREEDY algorithm ensures that all remaining sets have their correct priorities at all times, but LAZY-GREEDY updates the priority of a candidate set if and only if the set fails the size check.

5. Experimental results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>EAGER-GREEDY</th>
<th>LAZY-GREEDY</th>
<th>DF-GREEDY</th>
</tr>
</thead>
<tbody>
<tr>
<td>RETAIL size</td>
<td>181</td>
<td>180</td>
<td>185</td>
</tr>
<tr>
<td>RUNNING</td>
<td>406,372</td>
<td>406,406</td>
<td>406,428</td>
</tr>
<tr>
<td>ACCIDENTS</td>
<td>5,106</td>
<td>5,113</td>
<td>5,119</td>
</tr>
<tr>
<td>WEBDOCS</td>
<td>406,372</td>
<td>406,406</td>
<td>406,428</td>
</tr>
<tr>
<td>Running time</td>
<td>0.39</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>ACCIDENTS</td>
<td>1.53</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>WEBDOCS</td>
<td>73.10</td>
<td>5.34</td>
<td>5.11</td>
</tr>
</tbody>
</table>

Table 2: The smallest size and fastest time are bolded. The key parameter $p$ of DF-GREEDY is 1.1. The three implementations were executed on a multi-processor server with Linux OS and 32GB RAM.

Table 2 compares the two in-memory implementations and another implementation, DF-GREEDY, of a disk-oriented SET-COVER algorithm [4] that further approximates the solution quality. DF-GREEDY splits the sets into multiple buckets of different size ranges, and processes the buckets from the one with the largest range downwards.

The EAGER-GREEDY algorithm runs slower than the other two algorithms. The LAZY-GREEDY algorithm performs competitively against the DF-GREEDY algorithm in terms of running time and solution size.

6. Conclusion

The new algorithm, LAZY-GREEDY, avoids the inefficiency of EAGER-GREEDY by delaying set updates for as long as possible. The trade-off is that the worst-case running time is increased to $O(M^{3/2})$, compared to $O(M)$ for EAGER-GREEDY where $M$ is the total size of the input [3].
Memory Efficient Local Outlier Detection in Data Streams

Mahsa Salehi, Christopher Leckie, and Tharshan Vaithianathan
msalehi@student.unimelb.edu.au

Today, huge streams of data are generated, raising the challenge of how to detect outliers in data streams within limited memory.

Introduction

Outlier detection is an important task in data mining with applications ranging from intrusion detection to human activity recognition. With the growing need to analyse high-speed data streams, the task of outlier detection becomes even more challenging. Traditional outlier detection techniques cannot always assume that all the data can be stored for processing. Local outlier factor (LOF) is a well-known outlier detection algorithm that finds local outliers in data sets with varying data densities. LOF is an incremental version of LOF that assumes that the entire history of previous data points is available for the computation of LOF, and hence effectively requires unbounded memory. Here, we address this problem by proposing a memory-efficient incremental outlier detection algorithm, which has an accuracy close to Incremental LOF within a fixed memory bound. Our experimental results on synthetic data show that our approach has better memory and time complexity than LOF while having comparable accuracy.

Materials and methods

- **Bucket**: our memory limit
- **Clustering Model**: depends on the dataset; we used k-means here
- We store only one merged clustering: Updated Model plus bucket 2 last observations

Results

- Graph 1: Data point variance
- Graph 2: Data point number
- Graph 3: Time (s)

Literature cited


Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.
Preserving Location Privacy in Location-based Services
Obfuscation through Locality Preserving Hashing
Maryam Fanaeezpour, Lars Kulik, Egemen Tanin
Department of Computing and Information Systems
maryam.fanaeezpour@unimelb.edu.au

“Less than 90 days after government watchdogs reported drivers are not being clearly informed about privacy risks posed by vehicle-location data, the head of the U.S. Senate’s privacy committee announced legislation to give consumers more control over how information about their whereabouts is collected and used.”

How to provide private location-based queries?

**INTRODUCTION**
- High demand of using location-based technologies; GPS, RFID, GSM networks.
- Location-based queries; users ask queries about points of interest (POIs) in their proximity.
- Where is my nearest medical care center?

Users want to get their service on time without having to reveal their real location to an untrusted service provider:
- Range query: find POIs in user’s defined region.
- Nearest neighbor query: pinpoint the nearest POI in user’s vicinity.

**EXISTING METHODS AND CHALLENGES**
- Cryptographic techniques: not practical
- Transformation/obfuscation techniques
  - Locality preserving hashing techniques, in which points are transformed from high-dimension (2D) to 1D.
  - Space filling curves (SFCs) and locality sensitive hashing (LSH) are two subsets of this category.
  - Problem with applying SFC-based transformation: they can not guarantee the accuracy of the result because of the big jump in their traversing method.

**How to partition and index the spatial space effectively?**
- Providing quality of service: data quality (accuracy) and performance (speed, communication cost)
- Perserving location privacy

**OUR APPROACH: LOCALITY SENSITIVE HASHING**
- Two-layer partitioning scheme based on LSH concept to improve the level of accuracy for the NN queries compared to SFC-based technique:
  - Partitioning the space based on grid/quadtree; traversing it using SFC.
  - Superimposing the partitioned layer with Voronoi diagram of POIs to group the points.
  - Indexing the space and grouping the close points in a Voronoi cell into grid/quadtree partitions.
- **PRIVATE KEY:** various ways to traverse a partitioned space, using SFC, to create different hashed POIs tables on LBSs.

![Diagram](image)

**Table of values for grid/partitioned cells using LSH method**

<table>
<thead>
<tr>
<th>Index</th>
<th>Cell ID</th>
<th>Reference Value</th>
<th>Row Location</th>
<th>Column Location</th>
<th>Index of Hashed Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>k1</td>
<td>01</td>
<td>Hash(1)</td>
<td>a, b</td>
<td>00, 01</td>
<td>Hash(1), Hash(1)</td>
</tr>
<tr>
<td>k2</td>
<td>02</td>
<td>Hash(2)</td>
<td>a, b</td>
<td>00, 01</td>
<td>Hash(2), Hash(2)</td>
</tr>
<tr>
<td>k3</td>
<td>03</td>
<td>Hash(3)</td>
<td>a, b</td>
<td>00, 01</td>
<td>Hash(3), Hash(3)</td>
</tr>
<tr>
<td>k4</td>
<td>04</td>
<td>Hash(4)</td>
<td>a, b</td>
<td>00, 01</td>
<td>Hash(4), Hash(4)</td>
</tr>
</tbody>
</table>

**Different ways to go and index the space using SFC**

Superimposing a quadtree with partitioned space by Voronoi diagram of POIs.

73
Privacy Aware Dynamic Ride Sharing

Preeti Goel, Lars Kulik, Kotagiri Ramamohanarao

Introduction

**Dynamic Ride Sharing**

Instant Shared Rides

**Why?**
- Low Car Occupancy Rates
- Traffic congestion
- Emissions
- Travel costs (fuel and toll)
- Use of carpool lanes – faster
- Economical
- Reduces driver stress

Challenges

- Privacy
  - Driver / Passenger – The identity, time and exact location (route) information should only be disclosed at agreement or negotiation stage.
- Optimal Matching/ Cost
  - Passenger and driver time constraints
  - Shortest route with maximum revenue and lowest cost
  - Individual preferences

Our Approach

- Meeting points
- Driver - vehicle owner who wants to pick up people on his way provided he reaches his destination in a time limit.
- Passenger - A passenger is a user who wants a ride to reach a destination from a start meeting point (mps)

Driver Path Area Ellipse

Match Maker Model

Privacy Region Based Selection

Optimization

- Driver - optimal path computation - Trip cost, Trip duration
- Passenger - optimal driver selection - Trip cost, Start time, Trip duration

Experiments
In Bioinformatics, storage of genetic information has become fundamentally necessary for the future of the field. The read sequence alignments are stored in raw form, as plain ASCII text, which is useful for text-processing tools, but has the drawback of requiring more space than is necessary. We present a compressed representation for read sequence alignment data, with the goal of reducing the space for storing these data.

1. Read Sequence Alignment

Sequencing technologies produce millions of reads. Each read is a continuous fragment of data extracted from a genome, represented as a string of bases, letters than indicate the fundamental molecules of DNA. Later these reads are aligned, to identify regions of similarity that may indicate functional, structural, or evolutionary relationships between organisms.

2. Example of Data

3. Compression Encoding

A fixed-length encoding or variable-length prefix-free code is used to represent the symbols contained in the data, with the choice dependent on the probability distribution that arises over the set of symbols.

<table>
<thead>
<tr>
<th>Integer</th>
<th>Example Encoding Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Binary</td>
</tr>
<tr>
<td>2</td>
<td>010</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
</tr>
</tbody>
</table>

4. Method

Sequence are represented as independent components:

- A presumed sequence containing the bases that are more frequent in each overlapping position.
- Subsequences Not\(X\), with \(X \in \{A,C,G,T\}\), where Not\(X\) stores the read bases that differ with the presumed sequence, when the base in the presumed sequence is \(X\).
- A length array, that stores the length of each read.
- An offset array, that record the distance between reads.
- A copy array, storing the chunk lengths of the read that are equal to the presumed sequence.
- A replace array, containing the chunk lengths of the reads that differ with the presumed sequence.
- An overall list, IsN, of locations in the reads that are special cases (bases with N value).

Each independent component is coded using one of the representation for integers.

5. Results

Compressed size and encoding time (tested in a desktop computer Intel i7 CPU 3.40GHz):

<table>
<thead>
<tr>
<th>Alignment Data</th>
<th>NA12878</th>
<th>HG01477</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original (MB)</td>
<td>355.87</td>
<td>691.70</td>
</tr>
<tr>
<td>Bzip2 -9 (MB/sec)</td>
<td>61.85 / 36.96s</td>
<td>116.14 / 60.29s</td>
</tr>
<tr>
<td>Gzip -9 (MB/sec)</td>
<td>43.58 / 82.41s</td>
<td>79.87 / 153.58s</td>
</tr>
<tr>
<td>Ours (MB/sec)</td>
<td>35.24 / 39.16s</td>
<td>51.08 / 100.66s</td>
</tr>
</tbody>
</table>

6. References

SECURE REPAIR IN LARGE SCALE DISTRIBUTED STORAGE SYSTEMS
Lakshmi J Mohan ID: 656673 (MPhil leading to PhD)
Principal Supervisor: Assoc. Prof. Dr.Udaya Parampalli; Co-supervisor: Senior Lecturer Dr.Aaron Harwood

Large Scale Distributed Storage Systems
The widespread use of distributed applications necessitates very large scale distributed storage systems, making storage and back-up of data vital in today’s society. The massive volume of the data that needs to be stored poses significant challenges in terms of reliability and security.

To ensure reliability, a simple solution of replicating data three times was initially adopted by storage centres. But it became impractical as the data grew to the magnitude of 100’s of petabytes, due to the very large storage overhead involved. Erasure codes like Reed Solomon codes are now increasingly used in large scale distributed storage systems.

These provide high fault tolerance for small storage overhead. An (n,k) erasure code takes k data blocks and encodes it into (n-k) redundant blocks, such that any subset of k blocks (data or redundant or mixed) can reconstruct the original k data blocks. Nodes typically fail in distributed storage systems and failures are norm rather than the exception. When a node fails, the repair process has to download data from the surviving nodes and reconstruct the failed node.

If the system doesn’t experience failures and repairs, then we can store data securely in the DSS by mixing the data with a randomly generated key k. Security violation occurs when a node fails and is replaced by a new node.

Typical Repair Problem

Exact Repair tries to reconstruct the original lost data whereas functional repair tries to create a different block, maintaining the (n,k) fault tolerance of the code. The number of nodes contacted during repair(d) and the amount of data downloaded from each of them(b) are the primary factors that affect the repair network traffic or repair bandwidth.

Dimakis et al.[2] identified a trade-off that exists between the storage overhead and repair bandwidth needed for regeneration. Regenerating codes are a class of erasure codes that are optimal with respect to both storage space utilization and repair bandwidth. Minimum Storage Regenerating (MSR) codes minimize the storage whereas the Minimum Bandwidth Regenerating (MBR) codes minimize the repair bandwidth.

Secure Repair Challenge
Erasure codes naturally provide security by ensuring that an adversary obtaining access to less than k nodes can gain no information about the data. However, repairing can compromise the security of a distributed storage system.

There are two threat models:
Passive eavesdropper: The attacker gains access to a subset of the storage nodes. Active adversary: The attacker may corrupt the data stored in a subset of storage nodes and also pass erroneous data during repair operations.

OBJECTIVES
The outline of the research work is explained below:
1. Construction of secure optimum codes for distributed storage
2. Security analysis of the codes

The code constructed will be optimal with respect to both storage space utilization and the repair bandwidth. The code will be implemented in a real-time setup and its performance in terms of storage/bandwidth/repair requirements and computational complexity will be evaluated.

The performance of the code in the presence of adversaries will be studied and the implementation will be fine-tuned to be resistant to all types of adversarial errors.

RESULTS SO FAR
We worked with the library Jeraurke, which is an open-source library implementing the popular erasure codes.

Experiments were done on the distributed file system nCFS (Network Coding based Distributed File System), a proof-of-concept file system developed for distributed storage. The performance of RS codes in nCFS was studied along with the traditional RAID-5 and RAID-6 and another class of code called Exact-MBR code. The graphs show the download times of the various codes on the nCFS platform. The experiments were run in two set-ups, one is a 4 node setup through the University of Melbourne Ethernet and the other is a 6 nodes setup using a dedicated switch.

Results show that RS codes outperform RAID codes during failures and they could perform better as the number of storage nodes increase. Hence RS codes are still a popular choice in many practical storage systems including Facebook’s Hadoop HDFS.

The basics of erasure codes and the major types of erasure codes were studied.

We are now working with Hadoop framework, which is a more reliable and popular choice of erasure code implementation on Apache.

RELATED WORK
Spatial-Temporal Trajectory Simplification for Inferring Travel Paths

Hengfeng Li, Lars Kulik, Kotagiri Ramamohanarao
Department of Computing and Information Systems
The University of Melbourne
hengfeng.li@unimelb.edu.au

PROBLEM & CHALLENGES

(a) Ground truth
(b) Simplified trace

Problem:
How to map GPS traces to a road network accurately under noisy conditions?

Challenges:
- Noisy GPS points - Noise in GPS measurement causes the uncertainty of recorded locations;
- Stop points - If vehicles get stuck in traffic jams or have to stop at intersections, GPS points are still recorded;
- Variable road density - In some regions, roads are fairly well distributed. However, in urban areas, roads are densely placed due to the space limitations.

METHOD

We proposed three simplification algorithms to enhance map matching:
- Incremental Simplification (IS) simplifies a trajectory point-by-point by maintaining an incremental window;
- Sliding Window Simplification (SWS) keeps a fixed size of window moving forward with an increasing number of points;
- Global Simplification (GS) considers the entire trajectory while reducing the number of GPS points.

We use geometric property to determine the importance of a GPS point:
- Angular biased: \( f(s_1, s_2, \alpha) \rightarrow s_1 \cdot s_2 \cdot \alpha^3 \)
- L1 error norm: \( f(s_1, s_2) \rightarrow \sum_{s_1, s_2} \)
- Normalised linear: \( f(s_1, s_2, \alpha) \rightarrow (s_1 \cdot s_2 \cdot \alpha)/(s_1 + s_2) \)

RESULTS

We use following evaluation methods (ground truth \( P \) and predicted path \( P' \)):
- F1 Score Error Rate:
  \[
  \text{Precision} = \frac{\text{length}(P \cap P')}{\text{length}(P')}, \quad \text{Recall} = \frac{\text{length}(P \cap P')}{\text{length}(P)}
  \]
  \[
  F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
  \]
- HMM Error:
  \[
  d_- = \text{length}(P) - \text{length}(P \cap P'), \quad d_+ = \text{length}(P') - \text{length}(P \cap P')
  \]
  \[
  \text{HMM Error} = (d_- + d_+)/\text{length}(P)
  \]

CONTRIBUTIONS

- We propose three simplification algorithms - Incremental Simplification (IS), Sliding Window Simplification (SWS), and Global Simplification (GS);
- We use different weighting functions that incorporate spatial knowledge into the trajectory simplification process;
- We evaluate our algorithms on two real datasets - Seattle and Melbourne.
Transfer learning of a temporal bone performance model via anatomical feature registration
Yun Zhu, Ioanna Ioannou, Sudanthi Wijewickrema, James Bailey, Patom Pirrmchaisri, Gregor Kennedy, Stephen O’Leary
University of Melbourne

The aim of this work is to use transfer learning to adapt an surgical outcome assessment classifier built on a single temporal bone specimen to multiple specimens. Once a classifier is trained, we translate each new specimen into the original feature space, which allows us to carry out performance evaluation on different specimens using the same classifier.

Transfer learning framework

Step 1: We align the anatomical structures of the specimen model (i.e., the original specimen) with those of each new specimen using iterative closest points (ICP).

Step 2: Each voxel of the new specimen is transferred using the nearest anatomical structures and the classifier is applied to determine which voxels should be removed.

Step 2.2: The end-product is refined using surgical domain knowledge constraints.

Registration Result Example

The first graph shows the anatomical structures of specimen 6. In each of following graph, the two sets of magenta points represent the anatomical structure voxels in the specimen model. The set of green points represent the corresponding anatomical structure in specimen 6 before registration. The blue points represent the anatomical structure from specimen 6 after registration.

We acknowledge the support of IBES in the form of a PhD scholarship.
Thank You!

What a great time of planning it has been!

The 2014 CIS DC would not have been possible without the support and participation of students and staff at the Computer and Information Systems Department.

We would especially like to acknowledge Justin Zobel, the head of the CIS department and Rhonda Smithies for helping the CIS DC committee plan and co-ordinate the event.

Our dear supervisors who pushed us to submit to the colloquium and have supported us in our studies thus far.

Our sponsors, for making the CIS DC possible.

Our keynote speaker Dr. Conway, industry partners and colleagues from other institutions.

The panelists and judges at each session, who volunteered their time to make this event a success.

A heartfelt thank you to all students and staff who participated in the CIS DC.

Finally, we would like to acknowledge the 2014 planning committee for volunteering their time, skills and cheer, and also for the 2013 committee for their support.

The 2014 CIS DC Planning Committee!

Behnaz R. Yeganeh
Maryam Fanaeepour
Yi Han
Fernando Estrada
Yali Zhao
Alexander Kan
Ida Someh
Lida Rashidi
Kagonya Awori
Oliver Adams
A special thanks to our sponsors!