RESEARCH CONVERSAZIONE 2008
School of Information Technologies

Collection of Research Project Work
# TABLE OF CONTENTS

## Postgraduate Research Projects

<table>
<thead>
<tr>
<th>Page</th>
<th>Title of Poster</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identifying clinical concepts in a noise prone environment</td>
<td>Asgari, Pooyan</td>
</tr>
<tr>
<td>2</td>
<td>Automated Detection and Delineation of Lung Tumors in PET/CT Volumes Using a Lung Atlas and Iterative Mean-SUV Threshold Threshold</td>
<td>Ballangan, Cherry</td>
</tr>
<tr>
<td>3</td>
<td>Revolutionizing Information System Design</td>
<td>Budd, Peter</td>
</tr>
<tr>
<td>4</td>
<td>Direct Interaction with Large Displays through Low-cost Monocular Computer Vision</td>
<td>Cheng, Kelvin</td>
</tr>
<tr>
<td>5</td>
<td>My Files at a Touch: New Ways to Access Personal Information with Pervasive Interfaces</td>
<td>Collins, Anthony</td>
</tr>
<tr>
<td>6</td>
<td>Practical Fraud Detection for Business Using Record Linkage to gather valuable related information</td>
<td>De Vries, Timothy</td>
</tr>
<tr>
<td>7</td>
<td>On the Provisioning of QoS Mapping in Cellular and IP Networks using a Translation Matrix</td>
<td>Iftikhar, Mohsin</td>
</tr>
<tr>
<td>8</td>
<td>Provenance Data Management for Genome Sequencing</td>
<td>Jurnawan, Winly</td>
</tr>
<tr>
<td>9</td>
<td>Extracting ontological structure from collaborative tagging system</td>
<td>Lin, Huairen</td>
</tr>
<tr>
<td>10</td>
<td>Cophylogenetic analysis of emergent disease</td>
<td>Little, Robert</td>
</tr>
<tr>
<td>11</td>
<td>The information dynamics of phase changes in random Boolean networks</td>
<td>Lizier, Joseph</td>
</tr>
<tr>
<td>12</td>
<td>Efficient Channel Assignment Algorithms for Infrastructure WLANs</td>
<td>Manitpornsut, Suparerk</td>
</tr>
<tr>
<td>13</td>
<td>Advanced Remote Collaboration Grid</td>
<td>Ng, Choon Jin</td>
</tr>
<tr>
<td>14</td>
<td>Mining Student’s Learning Behaviour from Reflect</td>
<td>Rajibussalam</td>
</tr>
<tr>
<td>15</td>
<td>The health text pre-processing</td>
<td>Sabbagh, Mojtaba</td>
</tr>
<tr>
<td>16</td>
<td>Identifying interesting situations and modelling player behaviour within computer games</td>
<td>Storey, David</td>
</tr>
<tr>
<td>17</td>
<td>Complex Adaptive Service Systems</td>
<td>Wong, Waiho</td>
</tr>
<tr>
<td>18</td>
<td>Cross-Layer Hierarchical Mobility Management and Routing in Wireless Mesh Networks</td>
<td>Wu, Lei</td>
</tr>
<tr>
<td>Page</td>
<td>Title of Poster</td>
<td>Name</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------------------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>19</td>
<td>Answer Attenuation in Question Answering Systems</td>
<td>Bell, Katie</td>
</tr>
<tr>
<td>20</td>
<td>MetaView: Flexible views of users’ file collections</td>
<td>Bunton, James</td>
</tr>
<tr>
<td>21</td>
<td>Prototyper: a new filesystem access approach</td>
<td>Darke, Greg</td>
</tr>
<tr>
<td>22</td>
<td>Training Parsers using Redundancy of Information</td>
<td>Howlett, Susan</td>
</tr>
<tr>
<td>23</td>
<td>User-Defined Data Type for Genome Sequences</td>
<td>Huang, Zizhuo</td>
</tr>
<tr>
<td>24</td>
<td>Training a named entity tagger with Wikipedia</td>
<td>Notham, Joel</td>
</tr>
<tr>
<td>25</td>
<td>Flexible manipulation of sequence numbers in an interactive phylogeny analysis environment</td>
<td>Pang, Bo</td>
</tr>
<tr>
<td>26</td>
<td>Robust Snapshot Replication in a Database Cluster</td>
<td>Rodley, Mathew</td>
</tr>
<tr>
<td>27</td>
<td>Family: Documents, files, nodes and relationships</td>
<td>Thorogood, Sam</td>
</tr>
<tr>
<td>28</td>
<td>Real-time Multi-resolution 3D Line Rendering</td>
<td>Xu, Cong</td>
</tr>
<tr>
<td>29</td>
<td>Database Performance Evaluation Framework</td>
<td>Yoon, Danny</td>
</tr>
</tbody>
</table>
Note

This publication presents posters, and abstracts of research work which is to be shown on posters, which will be displayed at Research Conversazione 2008. All attempts are made to ensure that this publication truly reflects the actual status of the publications displayed at Research Conversazione 2008 at the School of Information Technologies. The material included in this publication is updated as at the time of publication and duplication of the material. However, if other material is subsequently added to the Research Conversazione event it may not appear in this publication. Further information on additions made after the cut-off date for this publication may be sought from the School.

Disclaimer

The School of Information Technologies accepts no responsibility for the content of the posters and abstracts herein published.

The material displayed in this publication, and at Research Conversazione 2008 is made available by students of the School for the benefit of industry and for comment and feedback from industry on work which in many cases is ongoing.
Identifying clinical concepts in a noise prone environment
Pooyan Asgari, School of Information Technologies
Health Information Technology Research Laboratory

Aims
- This research aims to introduce a new approach for identifying clinical terms and concepts within a noise prone collection of clinical notes.
- Introducing and evaluating a noise reduction strategy suitable for processing clinical notes.

Introduction
- The specific nature of a noisy data collection raises at least two major issues.
- The first issue comes from a scattered matrix of evidence for a specific concept in which however has many common thereby confounding attributes with other concepts.
- Considering more features or patterns with the hope of covering more rare situations may lead to the absorption of more noise by the system and impact the identification of other major terms and concepts, and therefore the overall performance of the system.
- The second issue comes from the nature of the data collection and the necessary process for gathering evidence about existence/absent of a specific concept.
- Assuming 4 possible answers for a search concept namely Exists/Not Exists /Questionable/Unmentioned, biases the decision algorithm towards the two more frequent classes: Unmentioned and Exists which have unlike characteristics.
- The Unmentioned label has to be identified based on lack of evidence for a given search concept while an Exists label should only be assigned in presence of clear indication of a given concept. Adding more features to the feature list in the machine learner leads the system to a better and more confident classification for Exists class but at same time may lead to inaccurate results for the Unmentioned answer due to an increase in the level of the noise.

Data
- The primary goal was to evaluate a proposed approach for processing a collection of 724 discharge summaries with a noise prone nature. The data contains possible evidence for at least 16 different diseases.

Methodology

![Diagram showing the methodology process]

Discussion
The proposed approach consisted of four major phases.
- In phase one, a deep negation analysis highlights the negative concepts within the free text and moves them to negation chest for later use. Moreover a rule based scanner , detects and removes misleading data sections within the text such as post surgical treatments or family histories.
- In Phase two, a minimal and natural feature set will be defined for all of 16 concept classifiers. We kept the number of features to monitor as low as possible based on the proposition that concepts are best defined in a few features and many features would add noise to the classifier. Features are restricted to those who meaningful and related to signs, syndromes and medication related to each concept.
- In Phase three, a subset of classifiers will be implemented. These are classifiers which their results can be used as a feature for other classifiers. Such as HTN (Hypertension) which is a risk factor for CHF (congestive heart failure). Based on the results of this classifiers, noise reduction rules and feature set for dependent classifiers will be adjusted.
- In final phase, using an external knowledge source, suspicious or irrational results will be filtered out.

Results
Evaluation has been done against given human performance as a gold standard with promising precision and recall of 0.969 and 0.969 respectively.

Future Work: dependent classifiers
- Dependent classifiers should be redesigned to take more advantages from higher level classifiers.
Automated Detection and Delineation of Lung Tumors in PET/CT Volumes Using a Lung Atlas and Iterative Mean-SUV Threshold

Cherry Ballangan, cherry@it.usyd.edu.au
Supervisors: David Feng, Xiuying Wang

Biomedical & Multimedia Information Technology (BMIT) Research Group
School of Information Technologies

1. Aim of the Project
   - To detect and delineate lung tumors in PET/CT images autonomously.

2. Introduction
   - PET/CT is now accepted as the best method to stage non-small cell lung cancer (NSCLC).
   - Interpretation of PET/CT is not trivial, and automated tumor segmentation would be beneficial.
   - Challenges in automated tumor segmentation:
     - Intensity similarity of the tumors with surrounding organs, such as the mediastinum and chest wall.
     - Irregularity of tumor’s shape.
     - Tumor extension to surrounding organs.

3. Proposed Method
   - Step 1: SUV (Standardized Uptake Value) Conversion
     If $x$ is a voxel’s activity concentration (Bq/cc) in the PET volume, the SUV of $x$ can be calculated using the following equation:
     \[ SUV(x) = \frac{W x}{D} \] \[ W = \text{patient’s weight (g)} \]
     \[ D = \text{injected radionuclide activity (Bq)} \]
     \[ t_1 = \text{PET acquisition time} \]
     \[ t_2 = \text{radionuclide injection time} \]
     \[ t_{1/2} = \text{radionuclide half life (s)} \]

   - Step 2: Nonrigid Lung Atlas Registration
     - Affine transformation
     - Spline interpolation

   - Step 3: Tumor Definition using Iterative Mean-SUV (IMS) Threshold
     SUV threshold can be calculated using iterative method:
     \[ \bar{SUV} = 0.307 \times M(\bar{SUV}_{n-1}) + 0.588 \]

   - Step 4: Tumor Delineation by Region Growing
     - Initial tumor in PET is used as a mask to achieve more accurate delineation in CT.
     - A seed is calculated from the centroid of the mask.

4. Experimental Results
   - We evaluated the proposed algorithm in 13 patients with non-small lung cancer (NSCLC) where FDG PET/CT scans were performed for staging.
   - Five additional thoracic CT volumes with healthy lungs were used to create the atlas.
   - All data are obtained from Royal Prince Alfred Hospital (RPAH) NSW.
   - The figures on the right show the experimental results over four patients with different tumor sizes, shapes, and locations.
   - The experiment results show that our method is able to tackle more complex cases such as when tumors are located adjacent to or involving the chest wall.

Figure explanation:
(a) : Fused PET/CT with the tumor indicated in square boxes
(b) : cropped images for PET/CT
(c) : cropped images for CT
(d) : tumor definition (fused with CT)
(e) : tumor delineation result (fused with CT)

5. Conclusion
   - We propose a new framework to detect and delineate lung tumors automatically by using an atlas and mean-SUV tumor definition. We use lung atlas to exclude the heart and liver to aid tumor.
   - Our initial experiments in 13 clinical datasets suggest that our method is promising.

6. Future Work
   - We will further evaluate the algorithm by comparing manual segmentation results from medical experts with our automated system.

Reference: SPIE Medical Imaging, 2009 (accepted for publication).
This research is supported by a BMIT research group ARC grant & MCIT/RI scholarship.
Clinical Information Systems

In the health domain, attempts to introduce clinical information systems have been going on since the 90’s with minimal success. The problems facing the implementation of clinical information systems however, are primarily societal and organisational, not technical. Since the 90’s, technology has improved markedly, but many of the original problems then are still problems now.

There are two major factors affecting the widespread implementation and use of clinical information systems in health IT, the rejection of systems by clinicians and the intractability of patient information once it is stored in a system. Both of these can be traced back to the limitations of the traditional methods of information system design.

User Rejection

No matter how effective an information system is, if it isn’t used and accepted by clinicians, then the system has failed. The reasons why clinicians reject systems are varied, but studies have discovered common factors:

- Knowledge acquisition rate too high
- Couldn’t see how it helped them
- No feedback for the effort put in by clinicians
- Reluctance to change work practices
- IT turnover made time investment unsustainable
- Focus on computer strength, not human lacking
- Input mechanisms clunky
- Forced changes to workflow instead of aiding
- Little understanding of how the information in the system would need to be used elsewhere

Data Intractability

Traditionally, the storage mechanisms of clinical information systems have been relational databases. The knowledge underpinning how the patient information is stored is implicitly defined and requires human involvement to understand. The sharing of that information between systems requires explicit mappings to be defined between data fields in information systems and is only accessible one way.

Most of the useful data in clinical information systems is in the form of natural language, or unstructured prose. Because traditional information systems have no way of dealing with text beyond storing it as a block, the semantic content of patient data is unusable for any computational purposes.

This intractability leads to serious issues in the health domain, where information in different systems is needed for patient care yet cannot be shared in any manageable way.

GCIMS

To overcome these problems, we have developed a framework for generating information systems at runtime from abstract definitions of information systems, called the Generative Clinical Information Management System (GCIMS).

The GCIMS project is based upon 3 pillars to overcome the limitations of traditional information system design: language as data, system interoperability and the Occam’s razor of design.

Language as Data

Because the majority of useful information in a patient note is in the form of free text, GCIMS uses Natural Language Processing (NLP) to extract semantic meaning from free text. This semantic extraction is used in every aspect of the GCIMS system, including interface design, where user input fields must be mapped to an external knowledge repository so that the semantic meaning of entered data can be preserved across information systems.

System Interoperability

Users want to feel like they are using their own system and customize it to their work practices so they have some feeling of ownership, which will reduce rejection rates. But each individual system must be able to share information with other systems. To achieve this, patient information must be separated from the manner in which it is used and become fully self descriptive, being tied to an external knowledge representation.

Once data has been separated, information systems need to be generated at runtime since the specific nature of the interfaces is not known til then.

Occam’s Razor of Design

In order to have a system that can generate information systems at runtime, we need to define the minimally complete set of system components that must be present. We used the Occam’s Razor heuristic maxim to produce the following components:

- Input and output interfaces
- Data storage
- Semantic data storage knowledge representations
- Workflow engine
- Management of the interface, knowledge and workflow
- Data analytics
Direct Interaction with Large Displays Through Low-Cost Monocular Computer Vision

Kelvin Cheng, kcheng@it.usyd.edu.au
Supervisor: Dr Masahiro Takatsuka
ViSLAB, School of Information Technologies

Introduction
- We investigate the feasibility of using monocular computer vision to allow bare-hand interaction with large display systems from a distance.

Pointing Models
- We studied the different pointing strategies and formalized three major models of targeting to describe models underpinning current hand pointing interactive systems.

\[ P_e = \text{eye position} \]
\[ P_o = \text{object/target position} \]
\[ l_g = \text{line of gaze from eye to object} \]
\[ P_p = \text{the pointing instrument} \]
\[ l_p = \text{line of pointing} \]

The Point Model

Our Interaction Model
Resultant on screen position would be user’s intended target

Dynamic Virtual Touch Screen
- Leverage advantage of large touch display
- Bring whole screen (virtually) towards user.
- Using dTouch at arm’s length
- User’s position is not fixed

Monocular Vision
- Eliminate computational cost for matching images
- Widely accessible to mass market
- Easy to setup and use
- However, presents challenge for finding depth information from a single camera

Monocular Positions estimation
- Geometric constrains in the environment were exploited
- Face detection was used to determine eye position - face width was used to determine the depth
- Fingertip position detected from lowest skin colour pixel
- The depth of the fingertip was calculated by intersecting a sphere (where the center is at the user’s shoulder, with arm’s length as the radius) and a line vector from the center of camera to the position of the detected fingertip in the camera’s image plane.

Naturalness and Accuracy
- Full arm stretch was found to be the most common pointing strategy (point or dTouch models)
- dTouch was found to be the most accurate strategy

Implementation
- OpenCV and Visual C++
- Logitech QuickCam Pro 4000
- Display size 81.5cm x 61.5cm
- Screen resolution 1024x768

Estimation Error
- Blue crosses show raw estimated positions for 5 seconds
- Red crosses show positions after Kalman filtering and user’s position assumed accurately estimated at 160cm

C – camera center
E – eye
F – fingertip
\( r \) – radius of circle
\( CF \) – vector from C to F

Natural and Direct – uses natural human ability and no prior knowledge necessary from users
Unintrusive – not wearing anything, less cumbersome, support multiple screens
Simple to setup and inexpensive - webcam
Unconstrained and untethered - view frustum changes as user moves – users can walk around while pointing
Remote pointing - used at a distance
Aims of the Research

- Tabletops are promising for supporting collaboration with digital information:
  - multiple people can sit face-to-face and interact with a shared workspace
  - Emerging pervasive interfaces, such as tabletops, create new challenges for interface design. How can we make it easy for people to do fundamental interaction tasks, such as accessing personal information?

- Tabletop hardware is appearing quickly (e.g., Microsoft Surface), but there is little research on how people bring and use real data at a tabletop — a critical aspect to explore if tabletops are to be used in practice.

Project Background

- File system access — a key facility for a personal computer — has not been researched for interactive tabletops.

- Pervasive media, such as tabletops, mark a new way to think about personal computing. The hierarchical file system has hardly evolved since the 1970s. In the advent of pervasive computing, we need to rethink fundamental facilities, such as file system access, and look to the future.

- We have explored and evaluated novel interfaces (such as Focus, in Figure 2) for supporting people collaborating with their files at a tabletop.

Aims of the Research

1. Gain understanding of the usability and usefulness of a conventional single-user file system interface for tabletop collaboration — should we just use Windows Explorer on a tabletop?

2. Compare with a new file system interface explicitly designed for multi-user collaboration to address the challenges of tabletop interaction (e.g. touch-based input).

Evaluating Collaboration with Tabletop File System Interfaces

- We designed a study to understand how conventional and novel file system interfaces are used for tabletop collaboration, to provide a foundation for future research. We compared:
  - Hierarchical access with Windows Explorer — a widely used and familiar interface for a large population of users (see Figure 1).
  - Associative access with Focus — a novel tabletop interface that supports multiple people collaborating with their files (see Figure 2).

Experiment Procedure

- We compared them on the key dimensions of Table use (e.g. clutter, table space use) and Collaboration (e.g. amount and styles of file interaction).

- 20 participants (in pairs) completed a holiday itinerary planning task that involved retrieving and discussing many files related to hotels and tourist attractions in two cities. They were assigned disjoint collections of files to organise beforehand.

- The tasks were repeated with each interface (with mixed ordering), but for a different holiday destination.

- Participants completed pre- and post-experiment questionnaires, a short interview, and the experiment videos were coded to obtain quantitative results.

Key Experiment Results

- A Wilcoxon signed-rank test indicates that participants interacted significantly more with their partner's files using Focus than with Explorer. (F mean = 30.2, E mean = 12.15, z = -3.07, p < .05)

- Participants took significantly longer to complete the tasks using Focus (F mean = 16.7, E mean = 8.8, z = -2.52, p < .05) but during this time participants discussed more files with Focus. (F mean = 30.2, E mean = 12.15, z = -2.88, p < .05)

- Significantly more touch conflicts (interruption of their partner’s actions) occurred with Focus than with Explorer. (F mean = 5.8, E mean = 1.3, z = -2.54, p < .05)

- Participants thought that their partner’s actions were more helpful for solving the task when using Focus (Likert-scale responses) (F median = 4.5 [SD=1.5], E median = 4 [SD=1.7], p < .05)

Evaluation Conclusions

- Using Explorer, a strong sense of file ownership made people reluctant to touch each other’s files, resulting in less discussion.

- While participants completed the tasks faster with Explorer, they discussed significantly more files during the tasks when using Focus — a potentially positive effect on collaboration.

- Explorer was usable at the tabletop, but had serious shortcomings (e.g. clutter and poor use of table space) that impacted on collaboration.

- The design of Focus encouraged file discussion, pointing to the importance of researching new file system interfaces for pervasive computing.

Contributions

- This work provides an important foundation for future research on small group file collaboration at tabletops.

- A critical feature for interactive tabletops is to allow people to bring, browse and use their digital files.

- We gained valuable insight into the issues and opportunities that arise when people conduct collaborative work using a shared workspace.

- Our results highlight the advantages of rethinking traditional file system interfaces and exploring new ones designed specifically for tabletop use.

Acknowledgements

This evaluation was conducted with collaborators from the HxI Initiative [braccetto] project at National ICT Australia.
Introduction

In order to design and build a functional fraud detection system we have to address issues which are often orthogonal to conventional data mining research methodologies consisting of classification, clustering, outlier detection and association rule mining. The main challenges we have to address are low quality data, the lack of training data and good features, and efficient runtime performance.

We detail a pragmatic approach to design a fraud detection system which has been implemented at a major insurance company. The novelty of our approach is the extensive emphasis on integrating domain knowledge through features and the use of a transparent logistic regression model for both data integration and fraud rating. The use of simple models leads to a more robust and reliable system. It also makes it easier to explain fraud ratings, false positives and other results to the system users and helps build confidence among domain experts, management and the legal department.

One major component of our system is the use of record linkage to link identity records over multiple insurance products. This approach provides the biggest source of new and interesting information for fraud detection purposes, as data sharing is frequently overlooked in most corporate situations where data mining is useful.

Key Requirements

Fraud detection has been critical in many large business areas for a long time, such as the stock market, e-commerce, credit card, retail telecommunication and insurance industries. Techniques to be used in this area can be rated using a number of important requirements:

- **Accuracy**
- **Efficiency**
- Providing an understandable result to the users
- Ease of training and tuning the model
- Implementation overhead
- **Ability for a low-cost ‘pilot’ to be created for testing**

Surveys of recent fraud detection techniques in the literature show that in nearly all cases, the focus is purely on designing an algorithm to have high accuracy. However, in real-world scenarios the chosen technique depends far more on the more practical issues listed above. Unsurprisingly, very few studies claim to have actual implementations, with most having a large emphasis on complex, non-linear supervised algorithms and using synthetic datasets. Some key studies found that in the long term, less complex and faster algorithms such as naive Bayes and logistic regression will produce equal or better results on data that suffices population and concept drifts, and changing adversarial threats. In this vein, our research focuses on the practical applications of fraud detection.

Identity Matching

Most large corporations today collect data at a significant rate, mostly for auditing purposes. However, these sources of information can be valuable when trying to solve a data mining problem. In particular, it is common the case that related data sources may be completely separate from each other. Sharing this data is often not a practice at large corporations, and carrying out data mining, to share this data often brings in the largest amount of interesting information.

Record linkage is a technique applicable anywhere where two related datasets have no shared database key. In this case, we can carry out an approximate or ‘fuzzy’ matching technique to link these datasets together.

In our implementation we use record linkage to link identities across two business systems, which provides a wealth of information related to the person at the center of the claim that is being investigated. Record linkage can also be utilised in a single database, to remove duplicates. In both cases, the aim is to find pairs of records that are highly similar. The use of identity matching in our system design is a critical requirement for any kind of complex data analysis where the behavior of the individual of interest and their associates have to be analysed. For example, identity matching can give us many extra claim records which belong to the same person as the claim being investigated, and similarly for the main claimant’s associates, causing on average more than a ten-fold increase in the amount of data available for analysis. The accuracy of the identity matching technique used is therefore crucial to avoid propagating error further down the system pipeline.

The most common approach for record linkage is the use of a classifier for comparing two records and usually assigning one of two classes, signifying that the comparison is a match or a non-match. The classifier requires a definition of how similar any two records are, which is usually solved with the use of a custom-built similarity metric. We created a set of custom similarity functions for each field containing information about the records being linked. Some examples that were used in our implementation are:

- First Name and Surname - The Jar–Winkler edit-distance function was used to estimate the number of typographical errors, if applicable, allowing for fuzzy matching. Phonetic encodings of names were incorporated to allow for pronounciation differences. The expected histogram shape and the areas in which these examples causes it to learn how to distinguish between the ambiguous and borderline cases with these examples.
- Date of Birth - We used a basic function to allow for some similar dates to pass through, allowing for common data entry errors or missing day information.

Results

These cases are naturally harder for both humans and the record linkage model to classify. Training the model with these examples causes it to learn how to distinguish between the ambiguous and borderline cases more rapidly than when trained on randomly selected cases. This method allows for a more accurate model for a given number of training examples, allowing the amount of effort required for hand-labeling to be minimised.

The expected histogram shape and the areas in which training data were selected for our experiments are shown in Figure 1. The expected threshold is shown near the middle of the two peaks, where most of the non-match comparisons are on the left of the threshold and most of the match comparisons are located on the right.

Blocking

The performance bottleneck in a system such as this will usually be the record linkage step, which has $O(n^2)$ unoptimised runtime requirements. Instead of comparing a new identity record against every other record in the database, we can apply a technique to discard sets of records that are unlikely to match. These are known as blocking or indexing methods, and are critical for vastly improving performance at the cost of a small amount of accuracy.

Fraud Model

Many techniques can be used at this stage, from outlier detection to standard classification, to detect various types of fraud. Our overall system design is shown in Figure 2.

Active Learning

Any comparison between two identity records can be used as an item of training data once it has been hand-labeled by a human. To generate this data, random selection can be made between records in the identity database. However, our experiments and past research found that selecting ambiguous examples for training purposes had a far greater effect on the model than for randomly selected examples. Ambiguous examples are generally cases near the borderline or threshold between the non-match and match classes.
On the Provisioning of QoS Mapping in Cellular and IP Networks using a Translation Matrix

Author: Mohsin Iftikhar, mohsinif@it.usyd.edu.au
Supervisors: Dr Bjorn Landfeldt, School of IT, University of Sydney
Dr Mine Caglar, Mathematics Department (Koc University, Turkey)

1. Introduction

- As mobile terminals move between access networks, it becomes an important issue how to map resource reservations between different networks in order to maintain QoS behavior.
- We present a novel mechanism for achieving service continuity through a Translation Matrix, which maps QoS parameters between different access networks.
- The matrices are used to compare existing flow parameters to possible new reservation classes as the terminal changes access network.

2. Approach

- It has been shown that aggregated traffic in 3G networks exhibit self-similar properties which are not captured by the current Poisson model.
- Our traffic model (self-similar and long-range dependent) belongs to a particular class of self-similar traffic models known as the telecom process, similar to a heavy tailed on/off process.
- We consider a G/M/1 queueing system, which considers multiple classes of self-similar traffic input and to build the translation matrix, we analyze the behaviour of the traffic passing through different kinds of DiffServ domains based on different queueing schemes.
- We present closed form expressions of packet delay and Packet Loss Rate for different self-similar traffic classes under Priority Queuing (PQ), Custom Queuing (CQ) and Low Latency Queuing (LLQ) scheduling schemes, which are currently used in commercial routers.
- We develop the finite queue Markov Chain for different scheduling disciplines and derive the corresponding transition probabilities.

3. Tools

We have developed a traffic generator for generating traffic according to our model. The generator can be used for experiments in multiservice mobile networks and is capable of generating traffic with a wide range of burstiness and heavy tailed compositions. We have also developed a comprehensive simulation tool for studies of G/M/1 systems under self-similar traffic that is readily extensible to any given scheduling scheme.

Together, the tools and models constitute a milestone on the way to proper traffic engineering of mobile networks.

4. Results: Simulation vs Test Bed

- This research is sponsored by ARC linkage grant DP0452162

5. Analytical Results: G/M/1 with PQ

\[
E[W] = \sum_{j=1}^{n} \sum_{k=1}^{m} \frac{1}{\mu_k} \prod_{i=1}^{k-1} \left( \frac{1}{\mu_i} + \frac{1}{\mu_i} \right) \cdot \frac{j_c(j_i, j_f, j_s, s_i) + j_c(j_f, j_s, s_i)}{\mu_k}
\]

6. Analytical Results: G/M/1 with LLQ

- The delay for Class 1 packets in the LLQ model is the same as for PQ, however for Class 2 and Class 3 packets, the exact bounds on packet delay are given as follows:

\[
C_3 = \sum_{j=1}^{n} \sum_{k=1}^{m} \prod_{i=1}^{k-1} \left( \frac{1}{\mu_i} + \frac{1}{\mu_i} \right) \cdot \frac{j_c(j_i, j_s, j_f, s_i) + j_c(j_i, j_f, s_i)}{\mu_k}
\]

7. Publications


8. Invited Talks

2. "Multiclass G/M/1 Queueing System with Self-Similar Input", Wurzburg University, Germany, 18th. Sep. 2006
Provenance Data Management for Genome Sequencing

Author: Winly Jurnanwan, wjur7699@usyd.edu.au
Supervisor: Uwe Röhm, roehm@it.usyd.edu.au
School of Information Technologies

Background
Modern Bioinformatics produces a huge volume of data daily. For example, the 1000 Genomes project produces 75TB of images and sequence data each week in one lab (Sanger Institute). A sequencing workflow consists of several steps, from chemical sample preparation, to image analysis, short read alignments and sequence assembly.

Manually recording and annotating every process step is a tedious work, and the rate of production would outperform the annotation process.

Therefore, automating the annotation of gene production process as provenance is utmost important, especially to spot the anomalies in the data or an error produce by the gene sequence production.

Case Study: 1000 Genomes Project
This is a collaborative project of human gene sequencing which is supported by several research institutions around the world. While it took 10 years and 3 billion dollars to sequence the first reference genome, the aim of this project is to collect high quality gene sequences of 1000 individuals in period of 3 years. The whole process is new and a new provenance system should be adopted [2].

Genome Sequencing Process of the 1000 Genomes Project

What is Provenance?
The word provenance initially is used in the field of historical studies to record the chain of ownership of an art object or a literature [1]. The word provenance is used interchangeably with the term lineage, pedigree, etc.

The same ideas has been implemented in the fields of e-science research. Where the provenance records serve as the information that confirm and authenticate the origin, value, history, owner and transformation that pertinent to a particular piece of data.

Provenance in E-Science
Provenance in computer science is metadata or annotation of data item's properties or workflow that has been applied to it [3]. Provenance in scientific research can further be classified as workflow provenance and data provenance which refer to information pertaining to processes that transform a data and information pertaining to the origin of the data respectively [4].

Provenance Application
There are many implementations of provenance in computer science to support scientific research area such as, bioinformatics, Biology, GIS, Chemical science, Earth science, Physic, Astronomy, etc. These implementations mostly focus on the area such as support on reproducibility, tracking the source of data, audit trail, error tracking and verification of data quality.

Aim of The Research
• Efficient management of provenance data of the 1000 Genomes project, from the wet lab sample preparation to the gene sequence assembly
• Aims of this research are:
  • Effective Provenance
    Analyse and indentified the crucial provenance information that should be recorded in 1000 Genomes project
  • Flexible Storage
    Store the provenance information in the RDBMS with extensible schema and query friendly manner.
  • Problem-solving support
    Allow the efficient query of provenance information for problem back tracking purpose
  • Virtual field
    Virtual field supports the flexible database schema extension, where the fields which is not available before can be added without affecting the database uptime. These new fields could be link back to the original data and act as the fields extensions.

References
1. Aims
- To derive ontological structure from folksonomies;
- To facilitate collaborative ontology construction and refinement.

2. Introduction
Collaborative tagging system has become a rapid growing web2.0 practice for annotating electronic objects.
As the amount of information annotated using tag has increased significantly, it has aggravated the difficulty of retrieving and exploring the annotated resources.
Significant research progresses in the field of semantic techniques have offered a promising prospect to overcome these problems.
The central aspect of the approach is the extraction of ontological structure, for the content of the web resources.

3. Background

Folksonomy
- The term of folksonomy was first proposed by Thomas Vander[1] to describe this kind of informal social classification structure where user use their own language/terminology to describe contents in collaborative tagging system.
- folksonomy shows a flat and non-hierarchical structure with unsupervised vocabularies

Ontology
- Ontology is the enabling technology for semantic web, providing not only clear and exact concept within a particular domain but also the relation between those concepts.

4. Existing problems
- Non-standard words are frequently used to quickly express user’s idea in folksonomy, which are often neglected in ontology building.
- The traditional ontology construction is a time consuming process, which need the attendance of the domain expert and the ontology engineering professionals.
- Most ontologies are out of date and cannot reflect the fast changing environment, including the knowledge changes and the usage changes.

5. Research Questions
- How to find pairs of high correlated tags from large dataset?
- How to construct hierarchical structure from these essential vocabularies – high correlated tags?
- How to involve the non-standard words in the hierarchical structure?
- How can ontology be evolved to reflect the fast changing environment, including the knowledge changes and the usage changes?

6. Research overview
1) Using low support association rule mining to find the pairs of high correlated tags – essential vocabularies for ontology building.
2) With the help of upper-level ontology, the standard vocabularies are organized into hierarchical structure by matching and mapping approaches.
3) Non-standard vocabularies, including compound and jargon are interwoven into the structure based on matchers using similarity and association rules.
4) Collaboratively maintaining the ontological structure into more formal and precise ontology.

7. Approach
- We have conducted preliminary experiment to generate high-confidence pairs of related tags with a low support threshold.
- We have built the hierarchy of related tags with ontology matchers using WordNet, similarity and association rules.
- Future works are focused on ontology evaluation and collaborative ontology evolving.

8. Preliminary results

Fig 2, a fragment of ontological structure in science domain

Fig 3. composition of vocabularies

References
Introduction

It is estimated that as many as 75% of emergent diseases are zoonotic, being transmitted from animals to humans [1]. Well-known examples of zoonotic diseases include avian influenza, BSE, HIV, SARS and Ebola hemorrhagic fever. Currently there is no statistical model of the evolution of zoonotic disease.

Cophylogeny attempts to uncover the relationships between distinct but dependent phylogenies (evolutionary trees). Typically the phylogenies in question are those of a group of dependent organisms, such as parasites and pathogens, and of a group of hosts. Cophylogenetic analysis determines the degree of codivergence, that is, how often the evolution of dependents matches that of hosts. Other evolutionary events of interest include duplication in which a parasite speciates in the absence of a host speciation, and lineage sorting, in which a parasite follows only one branch after a host speciation. One of the factors which makes cophylogeny a particularly difficult problem is the existence of host switches, events in which a dependent species "jumps" to a new host. It is very difficult to extract information about historical host switches from existing genetic and taxonomic data. Host switching is also precisely what gives rise to the emergence of zoonotic disease, and we therefore believe that a better understanding of host switching will lead to better methods of fighting or preventing the spread of such diseases.

Several approaches exist to the cophylogeny problem, primarily based on parsimony, attempting to explain the relationship between two phylogenies with the fewest non-codivergence events, or the lowest weighted event score. Others use tree metrics to measure the (dis)similarity of phylogenies; however a metric is by definition symmetric, while we generally assume that the evolution of hosts is dominated by environmental factors whereas the host itself constitutes the environment for the dependents, and hence that the relationship between the two is asymmetric. Our approach is cophylogeny mapping, a parsimony-based technique in which dependent evolutionary events are mapped onto specific locations in the host tree. We achieve this through the construction and analysis of jungles [3], specialized graphs which attempt to capture all plausible coevolutionary relationships between given phylogenies.

Objectives

This research is aimed at improving cophylogenetic analysis in several ways. We will extend an existing implementation of cophylogenetic analysis via jungles to deal with a wider variety of problems: reticulate phylogenies, in which a given species may have multiple parent species, allowing us to model the biological process of hybridization; multiple dependent phylogenies, allowing us to hypothesize evolutionary dependencies among the various parasites of a host; the use of estimated event dates to reduce the number of possible histories evaluated; and analysis of multi-host parasitism, relaxing the severe constraint that dependent species may only be found on a single host at any given time.

We will also develop a statistical model of cophylogeny by applying maximum likelihood estimation. The challenge here is the extreme paucity of available data, preventing the use of standard machine learning techniques. We aim to overcome this through the application of Markov chain Monte Carlo methods and Bayesian estimation. Such a statistical model will allow us to better evaluate the plausibility of a given solution to an instance of the cophylogeny problem.

Computational complexity

The general problem of reconstructing coevolutionary histories has recently been shown to be NP-complete [4]. If the relative times of all the evolutionary events in question are known, the problem can be solved in polynomial time. Unfortunately, this situation almost never occurs in practice, and if even some of the events are only known to have occurred at one of two possible times, the problem becomes NP complete (ibid). Woroby et al [8] illustrate the extreme difficulty of determining accurate dates even for very recent evolutionary events. Consequently we must rely on approximation and heuristics to render cophylogenetic analysis tractable.

Jungles

Jungles are specialized directed graphs designed to capture cophylogenetic relationships. In a jungle

\[ J = \{ H, P, \phi \} \]

where \( H \) is the host phylogeny, \( P \) the dependent phylogeny and \( \phi \) a map from the tips of \( J \) to the tips of \( H \), the nodes of \( J \) are labelled with dependent-host associations, while the arcs correspond to the evolutionary events hypothesized to lead from one such association to another. The triple \((H, P, \phi)\) can be visualized as a tanglegram (see figure 2).

We will extend the jungle-based approach to the case where dependents are associated with multiple hosts simultaneously by introducing a new event type, failure to diverge, in which a host lineage diverges and a dependent follows both new lineages without diverging itself. This involves determining all the ways in which failure to diverge can interact with the existing modelled events; in particular, host switches once again complicate matters vastly as there are many possible histories involving failure to diverge and switches for any non-trivial case of multi-host dependency.

Statistical modelling

Very little work has been done thus far on the statistics of cophylogeny, largely due to the absence of definitively solved problem instances to work from. Huelsenbeck et al [2] present a framework based on a simple statistical model of host switching. Their analysis uses similar techniques to those we will apply, but includes molecular sequence data, whereas we will work exclusively at the level of phylogenies. We will also incorporate a more sophisticated model of duplication and host switching, building on the work of Nee et al on random distributions associated with birth-death processes [5]. This will enable us to deal with cases in which several dependents occupy the same host lineage simultaneously, which the simpler existing model cannot.

References

The information dynamics of phase changes in random Boolean networks

Author: Joseph Lizier, jlizier@it.usyd.edu.au
Supervisor: Prof. Albert Zomaya; Dr. Mikhail Prokopenko (CSIRO)
School of Information Technologies

1. Aim
To study the phase transition in ordered-critical-chaotic behaviour in Random Boolean Networks (RBNs):
- as models of Gene Regulatory Networks (GRNs),
- from the perspective of distributed computation.

2. Random Boolean Networks
Discrete dynamical models of nodes [1]:
- In a directed network structure,
- which is determined at random from an average in-degree /K;
- With Boolean states,
- Updated synchronously in discrete time,
- using heterogeneous random update functions.

RBNs exhibit a well known phase transition from ordered-chaotic behaviour as a function of the connectivity /K:
(quantified with respect to δ - the Hamming distance between the final attractors of two initial networks states differentiated by only a single node’s state)

3. Why study RBNs?
- Their popularity as models of GRNs: network attractor represents cell type.
- Generality as network models with a large sample space – permits study of dynamics of networks rather than topology.
- Ideal platform for studying generalised phase transitions in networks.

4. Why study information dynamics in RBNs?
- Topology is well-established but time-series dynamics less understood.
- Much conjecture on phase transitions in information propagation and processing b/w ordered and chaotic regimes in RBNs;
- Our perspective of computation aligns with popularly held notions of information dynamics.
- Computation performed by the RBN to determine attractor has meaning as determination of cell type by GRNs.

5. Information dynamics
- **Information storage**: info in past of an agent relevant to predicting its future [2].
  \[ A_X(k) = I(X',X(k)) \]
- **Information transfer**: info provided by a source about destination’s next state that was not contained in the past of the destination.
  \[ T_{Y1,Y2}(k)=I(Y_{1,2};X^k) \]
- Complete transfer entropy also conditions on other causal information sources:
  \[ T_{Y1,Y2}(k)=I(Y_{1,2};X^k;Y^2) \]

6. Results and Conclusion

<table>
<thead>
<tr>
<th>Connectivity /K</th>
<th>Low ( /K &lt; 2 )</th>
<th>Mid ( /K = 2 )</th>
<th>High ( /K &gt; 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase</td>
<td>Ordered</td>
<td>Critical</td>
<td>Chaotic</td>
</tr>
<tr>
<td>Sensitivity to initial conditions</td>
<td>Low ( \delta &lt; 0 )</td>
<td>Critical ( \delta = 0 )</td>
<td>High ( \delta &gt; 0 )</td>
</tr>
<tr>
<td>Convergence of similar macro states</td>
<td>Strong</td>
<td>Uncertain</td>
<td>Highly divergent</td>
</tr>
</tbody>
</table>

1. Info storage maximised just on ordered side of critical regime.
2. (Coherent) info transfer maximised just on chaotic side of critical regime.
3. Balance near critical phase
4. Info modifications continues to increase into the chaotic regime.

References
1. C. Gershenson, ”Introduction to Random Boolean Networks,” Int. Conf. on Simulation and Synthesis of Artificial Life (AlifeIX), Boston, 2004.
1. Aims of the Project
- To improve the performance and stability of WLANs in dense deployment environments

2. Introduction
Channel assignment algorithms, e.g. [1-2], have been proposed in recent years, in order to minimize interference between neighbouring networks and hence improve performance. However, in very densely deployed and unplanned networks, it is impossible for current algorithms to assign interference-free channels to the wireless nodes as reported by [3]. In addition, they also induce the hidden node problem which leads to very poor network performance in terms of voice support.

3. Side-effect of traditional channel assignment algorithm
Weighted coloring-based channel assignment algorithm [1] (WCA) uses a weighted degree of interference (DOI) as objective function to minimize the overall radio interference. This approach causes the hidden node problem as a side-effect.

4. Hidden node problem in infrastructure WLANs

5. Our Approaches
- Weight of Affected Nodes by Hidden Nodes (WANH): Take bandwidth produced by hidden nodes (\(W_{\text{sum}}\)) into consideration!!!

Algorithm 1: WANH
1) Randomly choose channel \(c\) from a set of available channels and assign it to each base station.
2) Calculate \(P_{\text{WANH}} = \sum_{v \in \mathcal{V}} W_{\text{sum}}(v, c)\)
3) \(P_{\text{WANH}}(c) = \frac{1}{\sum_{c \in \mathcal{C}} P_{\text{WANH}}(c)}\)
4) \(P_{\text{WANH}}(c) = \frac{1}{\sum_{c \in \mathcal{C}} P_{\text{WANH}}(c)}\)
5) Randomize a real number \(a\), \(0.0 < a < 1.0\)
6) Choose \(c\) such that \(P_{\text{WANH}}(c)\) is the closest value to \(a\) and \(P_{\text{WANH}}(c) < a\)
7) Assign \(c\) to \(v_i\)
8) Repeat step 2 for all \(v_i \in \mathcal{V}\)

Note:
\[
W_{\text{sum}}(v, c) = W_f(c) + W_E(v, c)
\]
\[
W_f(c) = \sum_{v \in \mathcal{V}} |H(c)| \cdot H(c, v, c)
\]
\[
W_E(v, c) = \sum_{v \in \mathcal{V}} |H(c, v, c)| \cdot H(c, v, c)
\]
\[
H(c, v, c) = \begin{cases} 
1 & \text{if } v \notin L(c) \\
0 & \text{if } v \notin L(c)
\end{cases}
\]
\[
H(c, v, c) = \begin{cases} 
1 & \text{if } v \notin N(c) \\
0 & \text{if } v \notin N(c)
\end{cases}
\]

6. Results

7. Conclusions
We show that traditional channel assignment algorithms, e.g. WCA, can introduce higher hidden node probability than randomly assigning channels to wireless networks. In addition, we have shown that the RTS/CTS mechanism cannot completely alleviate the hidden node problem in infrastructure WLANs. We have further shown that the hidden node problem should be considered by channel assignment algorithms for WLANs since delay sensitive applications may suffer otherwise. Our proposed algorithms have been shown to improve WLAN performance significantly.

References

Acknowledgment: Department of Computer Engineering and Multimedia, School of Engineering, University of the Thai Chamber of Commerce, Thailand.
1. Aims of the Project
   To provide an architecture that can:
   • Unify multimedia protocols for the purpose of Computer Supported Cooperative Work (CSCW).
   • Extensible for future media/protocol.
   • Scalability in participants, services, and the type of media used.

2. Introduction
   Many collaboration solutions do not provide a dynamic collaboration environment. The definition of dynamic by our term is to share any media and application/services at anytime.

   Hence, the focus of this research is to allow the construction of a customisable collaboration environment which can be applied to any domain.

3. The Advanced Remote Collaboration Grid
   ![ARCG Features](image)

   **Figure 1:** ARCG Features.

   **Purpose:**
   The Advanced Remote Collaboration Grid (ARCG) provides an architecture that merges multiple services (through network streams) into the client. With these streams, participants can collaborate by feeding input back to the services, which manipulates the streams.

   **Benefits:**
   • **Does not make assumption on the types of stream:** It is the duty of the clients to determine the stream capability and decode them accordingly. Once decoded, the streams can be used to manipulate the client’s screen, and various other I/O interface such as audio and file system.
   • **Peer-to-peer (P2P) provides scalability:** At the network level, services and client can be located in the same machine. This forms a P2P network, as illustrated in Figure 2.

   ![ARCG Network Formation](image)

   **Figure 2:** ARCG network formation.

   • **Multicasting:** Since the type of streams is not assumed, it can range from a “heavy” stream such as video, to a “light” stream such as chat messages. The challenge remains in creating a scalable ARCG capable of coping with multiple “heavy” streams. To make this feasible, multicast will be widely utilised in this architecture.

   • **Service-based:** Decoding capability can be decoded on the fly. Moreover, service/client discovery must be seamless.

   The work of ARCG is still on-going and it can be separated into multiple stages described below.

4. SORC
   ![SORC Architecture](image)

   **Figure 3:** The SORC architecture.

   **Purpose:**
   The first stage of ARCG examines the benefits of a service based remote collaboration. With the Service-Oriented Remote Collaboration (SORC) framework, collaborative services are placed remotely. These services only exist in frame buffer format. Client sends feedback to the services which in turns manipulates its frame buffers.

   **How it works?:** The core element which makes all these possible is the proxy server sitting in between the clients and the server. It aggregates all services streams, combines and manages them into one stream before forwarding them. Client’s feedback are then rerouted to the appropriate service.

5. Reliable Multicast VNC
   ![Reliable Multicast VNC](image)

   **Figure 4:** The Reliable Multicast VNC (RMVNC).

   **Purpose:**
   One of ARCG goal is to support remote collaboration via remote frame buffering (RFB). However, current RFB solution for multicast are unreliable and do not support two-way interaction. The case for reliability is important especially when a WAN environment can easily cause data packet losses. Therefore, the goal of RMVNC is to enable the aforementioned features. This is an important step towards creating a scalable collaborative infrastructure.

   Moreover, RMVNC provides an initial testing ground on the type of reliable multicast protocol which can be used as the basis for the ARCG.

   **How it works?:** Additional protocol were defined to tackle late-join situation. If a client joins a session that has already started, then the protocol provide a mean for state transfer.

6. Ongoing Work
   Previous work on SORC and RMVNC yielded significant results. These projects forms the basis and provide guidelines on the direction of ARCG.

   The next step in ARCG is a major one:
   • To provide a foundation and infrastructure to integrate all different streams.
   • To devise a system and protocols for signalling between clients and services.
   • To integrate SORC and RMVNC so that RMVNC will be one of the service providing RFB sharing over the ARCG architecture.

---

**Advanced Remote Collaboration Grid**

**Author:** Choon Jin Ng, cjng@vislab.usyd.edu.au

**Supervisor:** Dr Masahiro Takatsuka

**School of Information Technologies**
1. Aims of the Project

The aim of this project is to explore data mining techniques to analyze complex educational data generated from an e-learning system called Reflect. The data mining aims to identify hidden information and patterns.

2. Introduction

Reflect has been used for many years at School of Information Technology University of Sydney to help students learning C programming. As a result, a large amount of complex data has been generated. We need powerful data analysis tools such as data mining (DM) to gain insights into useful patterns. Data mining has already been successfully applied in e-commerce but has limited applications to the educational data. One of the challenges to apply DM algorithms to educational data is that most DM algorithms have been designed to be used in e-commerce that has different purposes and orientations. Some of the differences are:

- Data: data produced by an e-learning system is more complex, sparse, dynamic and heterogeneous because it involves various user’s interactions and preferences while in e-commerce, it generally uses a simple web server log.
- Objective: the objective of implementation of DM in the e-commerce is to increase profit that is measurable for example by the amount of money earned and the increase number of customers while in e-learning the objective is to improve learning.
- Most DM tools such as DBMiner, Clementine, Intelligent Miner, etc. are not specifically designed for use with educational data and become difficult to use if an educator does not have extensive knowledge in DM.

3. Motivation

A large amount of data has been collected from the use of Reflect. It would be interesting and challenging to mine this data to search for interesting learning patterns within the data. This information has potential to improve learning and practices for students who will take similar courses and for educators who design the courses.

4. Research Questions

Through this research, I would like to answer these research questions:

- Can DM help to discover useful information that can be used to assist teachers in making a decision related to pedagogical aspects in designing a course?

5. Research and Methodology

In this research we have chosen to use two data mining tools namely Weka, a general DM software system distributed as a free open source software, and TADA-Ed (Tool for Advance Data Analysis for Education). One advantage of TADA-Ed is the tool has a filtering and preprocessing facilities in addition to visualization and DM algorithms.

Data collection.

The research uses the data from log of student interaction with Reflect, practical mark and exam results for the course of SOFT2139 in the year of 2007.

Data Preprocessing

Data is cleaned and transformed into an appropriate format and inputted into a database. After data cleansing, the number of instance available is 175 instances.

It is necessary to create a summarization table to summarize the information at required level (e.g. student) as the information about student’s interaction with the Reflect system are spread over several tables. For this purpose a mark_sum_TB has been create. We also create data discretization table as data discretization can increase the interpretation and comprehensibility.

Data summarization

The mark_sum_TB table has six attributes that integrates the most important information for our objectives as listed below:

- exam_mark : final exam mark
- tot_percen_lab : total percentage of lab work
- tot_percen_qz : total percentage of quiz
- final_prac_mark : final prac mark
- n_task : number of task student has done
- n_lo : number of learning objectives student has done.

6. Results

The following are summaries of results from running K-mean algorithm by using Weka and TADA-Ed.

<table>
<thead>
<tr>
<th>Group</th>
<th>Exam Mark (avg)</th>
<th>Lab work and Quiz (avg)</th>
<th>Final prac mark (avg)</th>
<th>Number of Task and Ed (avg)</th>
<th>Number of Instances (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong Hard work</td>
<td>Highest (49)</td>
<td>Highest (&gt; 49% )</td>
<td>Highest (27)</td>
<td>Highest (&gt; 49%)</td>
<td>29 (19)</td>
</tr>
<tr>
<td>Moderate High</td>
<td>Medium (27)</td>
<td>Medium (21)</td>
<td>Medium (14 - 22)</td>
<td>Medium (6 - 16)</td>
<td>27 (17)</td>
</tr>
<tr>
<td>Moderate Low</td>
<td>Medium (49-50)</td>
<td>Medium (14 - 22)</td>
<td>Medium (6 - 16)</td>
<td>Medium (6 - 16)</td>
<td>57 (36)</td>
</tr>
<tr>
<td>Strong Low work</td>
<td>Lowest (36)</td>
<td>Lowest (10)</td>
<td>Lowest (6 - 16)</td>
<td>Lowest (6 - 16)</td>
<td>24 (16)</td>
</tr>
</tbody>
</table>

From the first result table, it is shown that by using K-mean clustering algorithms, the teachers may identify four groups of students based on their characteristics and behaviour in performing exam and practicing the tasks in the Reflect system. In addition the second table shows which task that each group performs better in comparison to other groups and indicated by the exam mark.

7. Conclusion

The results have shown some interesting patterns that has potential to be use to learn student’s learning behaviour from their interaction with Reflect. For example the teachers may use the results to investigate the weakest students. Why they had not done more task on a specific domain, how many learning objectives they have done for a specific task etc. Regarding to the first research question, the research has produced some results that students may use to reflect their learning behaviour for example from their learners activities, resources and learning tasks that already done and their success or activities and tasks that have been done by other success students. In addition the result may also be used by teachers to get more objective feedback for evaluation of the structure of the course contents and its effectiveness on learning process.
Aims
- Anonymising health records
- Cleaning noise from narrative notes

Introduction
Pre-processing is very important for each system that needs to process a stream of text data. The pre-processing phase usually consists of changing the format of the input data in order to be suitable for the later processing. In health texts there is protected health information (PHI) which must be scrubbed from the text before they are used by other processing systems. Also the noise of a texts which are actually misspelling, abbreviations, mnemonics, and neologisms should be corrected or verified.

Text Pre-Processing Overview
We divide the system into two sub-system:

1. De-identification: No clinical document can be used for research before anonymisation. In this phase, our efforts are to devise a system which removes any piece of information that may lead to identifying a patient, such as name, date of birth, ID numbers, address, etc. There are ambiguities in some items which need to be resolved and without having enough clues from the sentence, we are not able to identify them correctly.

2. Spelling corrector: Spelling checkers are widely used in text editors. If a spelling corrector is designed for unlimited variety of texts the accuracy of it is quite low. As in each discipline there are special terminologies, expressions and abbreviations and so gathering all of them into one system is impossible, thus we are designing a customised spelling corrector for local health texts. This system identifies unknown words and suggests the best candidate for them.

De-Identification
Figure 1. shows some of eighteen classes of PHI. In some previous work the name of doctors also were removed from the text to increase the level of privacy protection. The de-identification algorithm includes four stages. (figure 2)

Spelling Corrector
There are two major views of spelling correction. Firstly, correcting isolated words and second, context dependent error correction. In this field we are concerned with the second part. Generally each spelling corrector consists of two phases. In the first stage unknown words are detected and some candidates are suggested. The next stage is ranking candidates and selecting the best one. In auto correction method the system selects the highest ranked word.

Error Sources
The source of errors is an important factor in choosing the correcting algorithm. Figure 3. shows three stages which an idea is matched to a word.

An error in each stage will lead to a misspelt word. If the mistake happens in the typing phase, it is easier to resolve than deeper phases (Spelling, Verbalisation). As we are not sure about the exact source of errors, we hypothesise that all types of error may occur. Figure 4. presents the scheme for suggesting candidates. Three more popular methods have been introduced.
Motivation

- Technology in the real world is approaching the point where recording information about states and actions with high precision is becoming feasible.
- This presents the possibility to record every single action of an individual and the changing state of the environment that they are in.
- It will produce millions of discrete pieces of data, which may contain interesting and useful information.
- Processing this data and extracting that information is a major challenge.

Motivation for use of Gaming Environments

- Computer games are effectively microcosms of the real world.
- Game worlds have well defined constraints on the actions an individual can take, and well defined interactions between players and the game environment.
- Data obtained from computer games is highly reliable and very precise as it can be obtained directly from the game engine.

Quake 3: Arena as an Experimental Platform

- Quake 3 is a first person shooter with small levels, simple game mechanics, and fast paced game play.
- An average round of play in Quake 3 generates around 100,000 events in total, most of them movement-related.
- The graph to the upper right shows the events of one player in a round of play involving 2 human players.
  - The red blocks mark out the easily identifiable combat situations.
  - There are many more situations that are not as easily identified.
  - Example: Each Killed or Died event marks the end of a fight situation, however there can be several fights interspersed with other situations before a kill/death.
  - It demonstrates that actions are easy to track, yet interpreting what is going on is a challenge.
- The image to the right was taken from a game involving 15 computer players.
  - Fighting is taking place on both ledges and in the lower area, involving individuals or groups.

Identifying and Examining Situations

- Situations are a period of time in which the player is engaged in something important to them, and is thus interesting to examine.
- Situations are represented as blocks of events.
  - Events are essentially changes in state and represent the fact that something happened.
  - Situations can be identified by either the occurrence of absence of specific events or sequences of events.
  - With situations identified, player behaviour within them can be examined.
  - Is player behaviour consistent?
  - Does their behaviour positively or negatively affect their performance?
Challenges of a global service system...

Change is chaotic in an increasingly connected world. Minor events could cascade into global phenomena. We propose that managing change in this new environment be centred around a complex adaptive systems framework.

A global enterprise has the ability to pull ideas from anyone in the world connected to it: customers, partners, employees. In times where ‘innovation happens everywhere’, how does the network transform ideas into customer value?

Global work teams are now the norm, facing challenges that are beyond traditional management tools. We propose network-centric models of team synchronicity and control to bridge the separation of time and space.

Small failures in complex networks could bring the whole system down e.g. cascading power failures. We investigate how a global service network is to contain local failures and prevent them from spreading through the network.

Keys: adaptive, network

Research Contributions

Network Structure of a Service System
A service system is structured as “a dynamic configuration of resources (people, technology, organisations and shared information)” (1). The network configuration that connects these resources will, undoubtedly, affect their efficacy. We investigate the effects of network structure on the delivery performance of a service system using a simulation model. The first phase of study involves the sharing of knowledge between service agents. Service agents in the model are capable of learning from previous results. They are connected through a network which enables knowledge sharing. Simulations of different network topologies reveal some interesting (preliminary) results.

Network topologies:
- Unconnected
- Fully Connected
- Hub
- Ring
- Small World

Given a pool of tasks to complete:
- A single task took, on average, longest to fulfil in a small world configuration. A task is completed quickest under a hub network.
- Comparing time-to-completion of all tasks, the hub and small-world configurations showed similar performance, beating all other configurations.

Control and Synchronisation in Service Systems
We apply control systems theory to a model of global work teams. Members of a team could belong to different organisations, physically separated with time-zone differences. Clearly, bringing a distributed team together and leading it towards a goal is a challenge; the latter being an issue of influence and control. Advances in control theory in complex dynamical networks provide some insight. (2) gives an overview on application of synchronisation theory in real world scenarios. Recent work on the issue of control advance two questions: i) how much control and ii) where to apply it for maximum effect (3). Already there is related work in (4) where these principals are applied to military command and control systems.

Our contribution will be to advance the application of this theory to service systems. We also aim to relate synchronisation with team capability for innovation.

Stability in Service Systems
Complex supply networks are robust when failures are random yet fragile if these failures occur at specific nodes. These networks, typically scale-free, display a preference of connections to certain nodes i.e. hubs. Their failures may not only partition the network, thus disrupting supply, but could also cause an avalanche of failures in the system e.g. the US Northeast Power Blackout of 2003. Although rare, the magnitude of these events give good reason for their study. As service networks begin to scale globally, there is a need to prepare for such failures.

We investigate two approaches:
- i. building resilience into the network through its topology
- ii. containing failures through network partitioning.

Both approaches will impact service performance. We look towards an optimal balance between performance and stability.

References
Motivation

- Wireless mesh networks (WMNs) have emerged as a ubiquitous solution for providing flexible wireless access at a low cost, but currently there is no efficient routing and mobility management schemes for large scale wireless mesh networks.
- Traditional Hierarchical Mobility Management and routing could be used for wireless mesh networks but WMNs usually has an unstructured graph topology which makes existing methods unsuitable.
- Layer 2 routing could be used in WMNs for better access to the status information at lower layers but MAC address based routing is not at all scalable.
- Hard Handoff between APs causes major delay in a handoff process.
- A hierarchical cross-layer mobility management and routing scheme is essential for efficiency purposes.

Achievements

- A complete solution for hierarchical mobility management and routing in large scale wireless mesh networks.
- Developed a greedy approximation algorithm for the NP-hard anchor point location problem.
- A Hierarchical Gateway-Aware Cross-Layer Routing protocol for WMNs (HGC-WMR) which over performs traditional ad hoc routing protocols in large scale networks with constant routing overhead.
- We observed close to fixed network performance in mobility management using our proposed approach.
- A virtual WIFI extension in madwifi to enable 'make before break' handoff which gives close to seamless performance in layer 2 hard handoff for delay sensitive applications.
- A layer 2 multi-hop binding protocol to the anchor points for inter-subnet mobility management which minimizes the micro handoff delay by eliminating address configuration.
- Achieved fault tolerance and load balancing by covering each MP with at least 2 anchor points.

Aims

- Design a complete framework for hierarchical cross-layer mobility management and routing in wireless mesh networks.
- Solve the problem of anchor point selection and subnet formation in WMNs to make WMNs hierarchical.
- Design a cross-layer routing protocol for both inter-subnet and intra-subnet routing.
- Develop a Layer 2 multi-hop handoff protocol for hard handoff of mobile clients.
- Develop a solution to reduce the long delay in Layer 2 hard handoff between APs.

Approaches

- Hierarchy in WMNs is divided into two layers with anchor points and portal points operating on layer3 and in-subnet MPs operating on layer2.
- Anchor point selection is solved as a facility location problem where anchor points are the facilities and MPs are the locations.

$$\text{Minimize} \quad \sum_{n \in N} \left( \sum_{n \in N} x_{nj} + \sum_{n \in N} y_{nj} \right)$$

subject to

\[ \sum_{n \in N} x_{nj} \geq 1 \]  
\[ t_j - x_j \geq 0, \quad \forall i \in C, j \in C \]  
\[ x_j, y_{nj} \in [0,1] \]  
\[ s_j \in [0,1] \]

- Program is NP hard, thus we created a greedy approach.

\[ \text{Sort all the MPs in } \mathcal{P} \text{ by } f_j \]

\[ \text{For } (\mathcal{M} \text{ in } \mathcal{P}) \]

\[ \text{Remove } i \text{ from } \mathcal{P} \text{ and put in } \mathcal{M} \]

\[ \text{Sort } \mathcal{M} \text{ by } f_j \]

\[ \text{For } j \in \mathcal{M} \]

\[ L = \text{avg cost based on equation (5)} \]

\[ \text{ Remove } i \text{ from } \mathcal{P} \text{ and put in } \mathcal{M} \text{ for } j \notin \mathcal{M} \]

\[ \text{ Reset } \]

\[ \text{ Remove } i \text{ from } \mathcal{P} \text{ and put in } \mathcal{M} \]

\[ \text{ End } \]

\[ \text{ End } \]

\[ \text{ End } \]
Answer Attenuation in Question Answering Systems

Katie Bell (kbel5892@it.usyd.edu.au)
Supervisor: James Curran
School of Information Technologies

Abstract
Research in Question Answering (QA) has been dominated by the TREC methodology of black-box system evaluation. Unfortunately, this makes it difficult to evaluate the effectiveness of individual components. We have exhaustively identified answer locations within the AQUAINT corpus for a sample of TREC questions, in doing so we also analyse the approach of humans to answering questions. The resulting corpus allows us to track answer attenuation through a QA system. We use this corpus to evaluate the Pronto QA system [1].

Answer Attenuation
We can see QA as a search problem, narrowing the document collection down to a single answer.

Definition: Answer attenuation is the proportion of correct answer instances remaining in the search space after each stage of processing.

In order to trace this, we need an exhaustive list of correct answer locations for a set of questions. This corpus will:
1. Determine the effectiveness of each component
2. Provide a more reliable overall QA Evaluation
3. Read through each document and highlight each correct answer

Correct and Supported Answers
The hardest part of annotating answers is determining which phrases actually answer the question. Consider this question:

How big is our galaxy in diameter?

Which of these is a correct answer?

- Approximately 100,000 light years
- Billions of kilometers
- Wider than you can comprehend

What if it were part of a sentence?
The galaxy is approximately 100,000 light years across.

Most galaxies are billions of kilometers in diameter.

“It’s wider than you can comprehend” is what Dr. Theakston said when asked how big the galaxy is.

The annotators also had difficulty determining when all of the correct answers in the document collection had been found.

Results
The average agreement between annotators for each question was 31.2%. However, the annotators each answered a common question, When was the Triangle Shirtwaist Fire? Of the 20 annotators, 14 or more agreed on each correct answer and a maximum on 9 annotators agreed incorrectly. This indicates that given enough annotators a majority decision will be reliable.

Answer Attenuation through Pronto
As shown in Table 1, for question A, Pronto retained most of the answers until the answer extraction stage (matcher), when all but one answer is lost. This one correct answer reached the top 10 but the top answer was incorrect.

Table 1: Answer Attenuation for 3 questions. A: Who created "The Muppets"? B: How big is our galaxy in diameter? C: How hot is the core of the earth?

<table>
<thead>
<tr>
<th>System component</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Answers</td>
<td>26</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Document Retrieval</td>
<td>19</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>CCG Parser</td>
<td>14</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Boxer</td>
<td>14</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Matcher</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>10 best answers</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Final answer</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The second question, How big is our galaxy in diameter?, had only 4 answers in the collection, 2 of these were found in the document retrieval stage, yet one of those sentences failed to be parsed, and was hence lost in the document analysis stage. The system returned an incorrect answer: 25.5 m. Question C was answered correctly: 11,000 degrees.

Conclusion
We created a corpus which can be used to trace the answer attenuation in any QA system, and determined that the creation of a larger and more accurate dataset is feasible. The final dataset will be useful both for system component analysis and as a more reliable overall system evaluation.

References

Question Answering Systems
The generic structure defined by Hirschman et al. [2] provides a useful framework to draw parallels. Processing starts with preprocessing of the document collection; indexing to improve retrieval time. The question analysis stage determines the expected answer type and generates keywords to be used in retrieving documents in the document selection stage. The document analysis stage selects sections from the text which are considered likely to contain answers. Short phrases which match the expect answer type of the question are selected in the answer extraction stage, these answers are then ranked according to how well they answer the original question.

Question Answering Evaluation
There are three types of evaluation of Question Answering systems.

1. Human Judging The gold standard. By far the most accurate, but also the most expensive.
2. TREC Answer Keys An approximation of the TREC human judgements, using manually crafted regular expressions to match to the answers
3. Specific Components There are good evaluations specific to individual components, but are not generalisable to all components
4. All Components The only current methods are either manually done, or ablation experiments which reveal very little about the system.

Collecting the Corpus
We created a web-based interface for finding and annotating the answers to questions. A group of 20 volunteers spent 5 hours annotating, resulting in answers for 78 questions.

Document Collection: The AQUAINT-1 corpus of newswire text, approx 1 000 000 documents.

Questions: Taken from the 1999 and 2000 TREC QA Tracks.

Figure 1: A demonstration of Answer Attenuation over a generic QA system structure

Figure 2: Screen of the web-based annotation interface
MetaView: Flexible views of users’ file collections

James Bunton (jamesbunton@fastmail.fm)
Judy Kay & Bob Kummerfeld
School of Information Technologies

Features
Flexible views of file collections structured according to the user's current needs. Views can be restructured automatically whenever a user desires. The user's existing file system organisation is left untouched by MetaView.

System Overview
MetaView allows users to create views, which are populated with directories and files such that the path to a file describes its contents according to the metadata which the user is interested in. If we think of the full path name of a file in a conventional hierarchical file system as an ordered series of metadata tags, essentially MetaView makes it possible to generalise that to allow other orderings of the metadata to create different views. This is implemented based upon symbolic links on a regular file system which allows MetaView to retain full compatibility with all existing software and makes installation simple. The following figures are screenshots from the Mac OS X Finder, looking at various views as described below. For example, a user may decide to have several views of their music files. Let us suppose that the first is a flat list of files as shown in the small subset of a user's music files in Figure 1. This shows just the first few of a large number of music files. The second example shows a view allowing the user to browse by genre, then artist and finally MetaView makes it possible to generalise this to allow other orderings of the metadata to create different views.

Background
Hierarchical file systems are the most common way of organising large collections of documents. However, there are several desirable features they do not support. These include: good support for placing files in multiple locations; dynamic views on the users' data; and explicit ordering of files. MetaView allows users to describe how they wish to view their files by specifying an organisational structure based on a metadata path such as Genre/Artist/Title. Experiments indicate that this approach is viable for collections of up to several thousand files in size, enabling flexible organisation of substantial parts of a user's file system.

Scalability Evaluation
An evaluation was conducted with a 10GiB collection of music from opsound.org with 1836 MP3 files. For each of these files Spotlight allows access to metadata such as artist, title and usually multiple genres. The testing was performed on an Intel iMac running Mac OS X 10.5.5 with a Core 2 Duo 2.16 GHz processor and 1GiB of memory. MetaView took 20 seconds on average to construct a view in the format: Genre/Artist/Title.mp3. To update the view after changing one file took 3 seconds on average. Almost all of this time was spent in the stat() call, required to determine which file in the list from Spotlight was the changed one. Additional tests were performed with different sized file collections, these can be seen in Figure 3 and 4. All tests were performed 3 times while the system was under no load and the result was averaged. In all cases there was less than one second difference between the trials.

Further work
A save dialog allowing users to tag files, displaying a list of commonly used tags, would be a good way for users to apply custom metadata to their files. To address the performance issue mentioned in the evaluation MetaView could use the FSEvents API to watch for changes to files on the system, in this way it would know which files have been changed in a Spotlight result set without resorting to a stat() of each file. A better solution would be for Spotlight to provide updates to existing result sets rather than reposting the complete list of files each time. Support for multiple documents within a file, such as mbox files or compressed archives would be another useful feature. Complementing this would be support for "bundles", or documents that are spread across multiple files, common with Mac OS X applications.

Conclusion
MetaView provides a flexible and adaptable new means to organise and access files. While it is consistent with the mental model of traditional hierarchical file systems that most users are familiar with today, it empowers the user to create arbitrary new organisation schemas, while preserving compatibility with existing software and work practices.
1 Problem
We introduce the notion of a digital artefact to be defined as anything a user may keep on their computer and think of as a distinct entity, for example: Emails, Word processor documents

There are many times when a person works on an activity that makes use of many digital artefacts which are spread across different parts of the their filesystem.

Currently it is tedious for people to assemble these because they are in different parts of the filesystem. This project aims to create a new tool called Prototyper to organise digital artefacts according to activities.

Why this is a problem?
One of the fundamental facilities provided by operating systems is a way to organise files into groupings using folders. This provides the predominant organisational structure for the digital artefacts which are represented by these files. In widely used operating systems, this organisation is hierarchical. Although this works well in many cases, it does have several severe limitations, one of which is that a digital artefact can only be stored in a single location.

There have been attempts at addressing this problem with hierarchical systems such as symbolic/hard links under Unix and shortcuts within Windows. This, however, is only a partial solution; while it works reasonably for expert users, it is used little by the general population.

One widely used alternative way of accessing files is to search, using tools such as Spotlight [1] and Desktop Search [3].

Example
Consider a hypothetical user, Jan Smith, an office worker who has been asked by her manager to give a progress report about an upcoming product to the industry. She will base this report from previous one. Table 1 shows some of the activities that are on Jan’s system.

<table>
<thead>
<tr>
<th>Type</th>
<th>Metadata</th>
<th>Activity metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email</td>
<td>To: Jan Smith From: Manager Subject: Request for Q3 talk</td>
<td>A4</td>
</tr>
<tr>
<td>File</td>
<td>Name: Presentation-G1.ppt Folder: /Tasks</td>
<td>A3, A4</td>
</tr>
<tr>
<td>File</td>
<td>Name: Presentation-G2.ppt Folder: /Tasks</td>
<td>A4</td>
</tr>
<tr>
<td>File</td>
<td>Name: Presentations.pdf Folder: /just/share/guidelines</td>
<td>A3, A4</td>
</tr>
<tr>
<td>Email</td>
<td>To: Jan Smith From: Pat Brown Subject: Q3 Status Update</td>
<td>A3, A4</td>
</tr>
<tr>
<td>Email</td>
<td>To: Manager From: Jan Smith Subject: Confirmation of Q3 details</td>
<td>A4</td>
</tr>
<tr>
<td>File</td>
<td>Name: Q3 Executive report.doc Folder: /usr/share/reports/2008</td>
<td>A2, A4</td>
</tr>
<tr>
<td>File</td>
<td>Name: Q3 Full report.pdf Folder: /usr/share/reports/2008</td>
<td>A2, A4</td>
</tr>
<tr>
<td>File</td>
<td>Name: Q3 Budget.xls Folder: /reports/2008</td>
<td>A2, A4</td>
</tr>
<tr>
<td>File</td>
<td>Name: Q3 Project Timelines.pdf Folder: /Notes</td>
<td>A1, A2, A3, A4</td>
</tr>
</tbody>
</table>

Table 2: Jan Smith’s filesystem (digital artefacts related to A4)

2 Approach
We build upon the idea of grouping digital artefacts around activities from previous systems such as UMEA [4], and TaskTracer [2]. Users found that these systems added an unacceptable extra burden of manually tracking when they change activities.

A key to our approach of automatically associating digital artefacts comes from prototype theory [5]. Essentially we ask the user to identify one digital artefact which they think defines the task. Prototyper uses this to find the other documents that are related to it. Then Prototyper aims to automatically identify the other digital artefacts associated with this activity.

3 System Design
Prototyper has three key parts; the reasoner which attempts to automatically associate files to activities, the event source which provides notification to the reasoner when files are in use, and the file metadata which keeps track of previous associations. This architecture can be seen in Figure 1.

5 Conclusion
Prototyper represents the first exploration of the use of prototype theory to derive a new approach to filesystem access based on automatically tracking use of digital artefacts to associate them with user’s activities.

Acknowledgements
This work was supported by the Apple University Consortium.

References

Figure 1: System design
Introduction

To enable more effective human–computer interaction, computers need a greater level of understanding of how we communicate with natural language. Current Information Retrieval systems like Google operate on bag-of-words or keyword approaches, which fail to capture the relationship between words. This is inadequate, since

\[ \text{Oswald shot Kennedy. is not the same as} \]

\[ \text{Kennedy shot Oswald.} \]

One step towards natural language understanding is the analysis of syntactic structure, or parsing.

Figure 1: Alternative ways of representing syntactic structure and the difference between “Oswald shot Kennedy” and “Kennedy shot Oswald”.

For a computer to learn to parse sentences, it needs a large amount of annotated training data, however constructing this data by hand is time-consuming and expensive.

Aim

We want to automatically create annotated training data to supplement existing training data for a state-of-the-art parser (called the C&C parser [1]) by exploiting the redundancy of information on the Web.

Redundancy of information

Mozart was born in 1756.

Wolfgang Amadeus Mozart (baptized Johannes Chrysostomus Wolfgangus Theophilius) was born in Salzburg in 1756, the second survivor out of six children.

is more difficult. However, the words common to the two sentences are related in the same ways. The sheer size of the Web means the same piece of information will be repeated in a number of different contexts and many of these are likely to be simple sentences like the first, so we can identify the relationships with a high level of confidence and apply them to the complex sentences.

Collecting sentences

Not all words will be consistently related across sentences, so we use facts consisting of relatively unambiguous keywords, such as

\[ \text{Mozart was born in 1756.} \]

\[ \text{Jennifer Capriati is a tennis player} \]

\[ \text{Marie Curie discovered radium} \]

We use each set of fact keywords to collect HTML documents (A), then divide the documents into sentences and keep those sentences that

- consist of ASCII alphanumeric characters and a small number of other punctuation characters,
- have at most 40 tokens (roughly, words), and
- contain all of the fact keywords. (B)

Extracting constraints

We consider simple sentences to be those that

- are at most 20 tokens long, and
- contain exactly one finite verb (specifically, one word with part-of-speech tag VBD, VBP or VBJ). (C)

We use the C&C parser to parse the ten shortest such sentences for each fact (D) and extract the chain of relationships connecting the fact keywords (E). We accept relationship chains if they appear in two or more of these ten sentences.

Creating new training data

Finally, we take all the sentences collected for a fact and parse them with the C&C parser (G), placing constraints (F) to ensure that the analysis generates the desired chain of relationships. If a valid analysis is found, the sentence becomes part of the additional training data, annotated with the parser's analysis.

Training the parser

The model for the C&C parser is normally trained on sections 02–21 of CCGbank, a corpus of newswire text [2]. We train two separate models, one using only CCGbank (baseline model), and one using CCGbank plus the additional training data (new model). (H)

Future

Possible extensions to this work include the following:

- Perform an evaluation of the two parsing models on a test corpus of Web text, since genre effects are a factor in performance.
- Investigate whether data collected using different kinds of facts affect performance in different ways.
- Increase the scale of the data collection to test the limits of the technique, whether more data is better or if the level of noise is too high.
- Extend the implementation to allow constraints to be enforced in a wider range of circumstances.
- Examine the sets of constraints generated by different definitions of simple sentences.

Conclusion

Since the sentences added to the training data are automatically annotated, their annotations are potentially noisy. We have shown that despite this, parser performance has not suffered, and may even be slightly improved. As there is still a vast number of facts that could be used as input to this process, there is potential to increase our training corpora by a considerable amount without human intervention.

Results

Following [1], we evaluate the C&C parser by calculating labelled and unlabelled precision, recall and F-score over the dependencies in section 23 of CCGbank, along with sentence and category accuracy and coverage. (I)

Table 1: Performance of the two parsing models on CCGbank section 23. With the exception of LF (POS), results use gold standard POS tags.

References


A summary of this work will also appear in the Proceedings of the Australasian Language Technology Work shop 2008.
The World of Database:
- ACID property
- Transaction control
- Standardized data manipulation language

The World of Scientific Databases:
- Huge volume of data
- Data managed by highly efficient algorithms
- Most data stored in binary files or plain text

Research Question

The World of Scientific Databases

We propose 2 different ways of treating FASTQ records:

Designs

Two important things to point out here is that, the

Sequence attribute consist of only characters

A,C,G,T

and

Quality attribute is a series of integers mapped to

ASCII characters.

Designs

We propose 2 different ways of treating FASTQ records:

SingleFASTQ: where each row correspond to one single

FASTQ record.

BatchedFASTQ: where each row correspond to multiple

FASTQ records.

Experiment Results

The design was implemented in Microsoft SQL Server 2008.

As shown in Figure 6, Zip is the obvious winner in regard of data size, however we should keep in mind that the retrieval of any data from Zip implies the need to decompress the whole file. Although SingleFASTQ loses to Zip, it does out perform the two compression methods provided by SQL Server 2008.

Figure 7 shows how batch size effects the data size for BatchedFASTQ data type:

Figure 8 compares the original file against Zip, SingleFASTQ and the best performing BatchedFASTQ.

Conclusion

Our best performing approach BatchedFASTQ with a batch size of 500 has a data size of 344.469 MB which is relatively close to the 222.8 MB achieved by Zip, and not having the need to decompress all data before retrieval. We are confident that the result of this research can aid in bridging the gap between modern DBMSs and specialized bioinformatic systems.

Acknowledgements

- Supervisor: Dr.Uwe Roehm
- SIT Database Research Group
- Microsoft Corporation, Redmond
- The Wellcome Trust Sanger Institute, UK

References


Figure 6: Comparison of data size (in MB) between different approaches

Figure 7: Comparison of data size (in MB) between different batch sizes

Figure 8: Comparison of data size (in MB)
Named Entity Recognition
A central task within Natural Language Processing is Named Entity Recognition (NER), wherein a system identifies and classifies Named Entities (NEs) within text, such as names of people (PER), organisations (ORG) and locations (LOC).

Ambiguities and unfamiliar terms make this challenging, as in the following NE-tagged sentence:

[PER Paris Hilton] visited the [LOC Paris] [ORG Hilton].

Rules and lists may be used to solve many cases:
Mark capitalised words after [Mr., Dr., Gen.] as PER
But rule construction can be difficult, and does not easily adapt to new domains or languages.

Statistical taggers are more flexible—they learn common entity patterns—but they rely on costly, expert-annotated training corpora. NE-annotated corpora are generally small. Compare:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC-7</td>
<td>83081</td>
</tr>
<tr>
<td>CoNLL-03</td>
<td>209208</td>
</tr>
<tr>
<td>BBN</td>
<td>910184</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>293,821,386</td>
</tr>
</tbody>
</table>

Table 1: Annotated corpus sizes

Wikipedia: free training data
Wikipedia [1] is an enormous, free encyclopedia. We have automatically transformed it into an annotated corpus for NER training, by (see Figure 1):
1. Classifying all Wikipedia articles into entity classes
2. Splitting Wikipedia articles into sentences
3. Labelling NEs according to link targets
4. Selecting sentences for inclusion in a corpus

Classifying articles
We classify all Wikipedia articles into: NE classes; non-entity (NON); disambiguation page (DAB); or unknown. Heuristics are learnt initially from 1300 hand-labelled articles, and confident classifications provide new heuristics through bootstrapping. Features include:

Category head nouns: Plural phrasal head nouns of Wikipedia categories often indicate NE class.

The article Port Melbourne’s membership in category “Suburbs of Melbourne” identifies it as a LOC.

Definition head nouns: If an article’s first sentence includes the word ‘is/are/was/are/were, we assume the head of the noun phrase that follows defines the article topic and indicates its class.

Class-specific rules: Lowercase in-links indicate NON; article categories or title identify DAB.

Table 2 shows our state-of-the-art classification results:

<table>
<thead>
<tr>
<th>Category</th>
<th>MUC-7</th>
<th>CoNLL-03</th>
<th>BBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>77.9</td>
<td>84.1</td>
<td>81.5</td>
</tr>
<tr>
<td>CoNLL</td>
<td>84.1</td>
<td>89.2</td>
<td>84.1</td>
</tr>
<tr>
<td>BBN</td>
<td>81.5</td>
<td>84.1</td>
<td>84.1</td>
</tr>
<tr>
<td>All</td>
<td>81.5</td>
<td>84.1</td>
<td>84.1</td>
</tr>
</tbody>
</table>

Table 2: Classification results averaged over 10-fld cross-validation

Wikipedia articles:

- Shakespeare
- Hamlet
- Sydney
- Sacramento
- Paris
- France
- California
- Dr. Smith, Jones & Co.

Table 3: F-scores for various TRAIN and TEST corpus combinations with a 3.5-million word Arabic Wikipedia-derived corpus.

<table>
<thead>
<tr>
<th>Test corpus</th>
<th>TRAIN</th>
<th>MUC-7</th>
<th>CoNLL</th>
<th>BBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>66.9</td>
<td>69.3</td>
<td>67.5</td>
<td>65.4</td>
</tr>
<tr>
<td>CoNLL</td>
<td>81.2</td>
<td>82.8</td>
<td>82.1</td>
<td>62.4</td>
</tr>
<tr>
<td>BBN</td>
<td>54.7</td>
<td>86.7</td>
<td>77.9</td>
<td>53.9</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>60.9</td>
<td>69.3</td>
<td>67.5</td>
<td>65.4</td>
</tr>
</tbody>
</table>

Table 3: F-scores for various TRAIN and TEST corpus combinations with a 3.5-million word Arabic Wikipedia-derived corpus.

When training and test data have different sources, performance is generally poor. Often Wikipedia performs better than standard corpora do upon each other.

More data is better up to a point:

Fixing links
Wikipedia linking is much less systematic than NE annotation. Some links therefore need to be adjusted:

<table>
<thead>
<tr>
<th>Link</th>
<th>NE-Tagged sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peru Shakespeare’s Hamlet</td>
<td>[PER Shakespeare’s Hamlet] [MISC Hamlet] [LOC Sydney, Australia]</td>
</tr>
<tr>
<td>Peru Prime Minister Kevin Rudd</td>
<td>[PER Prime Minister Kevin Rudd]</td>
</tr>
</tbody>
</table>

Table 3 shows our state-of-the-art classification results:

<table>
<thead>
<tr>
<th>Category</th>
<th>MUC-7</th>
<th>CoNLL</th>
<th>BBN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUC</td>
<td>77.9</td>
<td>84.1</td>
<td>81.5</td>
</tr>
<tr>
<td>CoNLL</td>
<td>84.1</td>
<td>89.2</td>
<td>84.1</td>
</tr>
<tr>
<td>BBN</td>
<td>81.5</td>
<td>84.1</td>
<td>84.1</td>
</tr>
<tr>
<td>All</td>
<td>81.5</td>
<td>84.1</td>
<td>84.1</td>
</tr>
</tbody>
</table>

Table 2: Classification results averaged over 10-fld cross-validation

Concluding and future work
We show that Wikipedia is a clearly viable source of huge, free, up-to-date NER corpora. Its exploitation may still be refined for improved accuracy and coverage. We may also be able to better augment available training data by automatically selecting beneficial sentences, or using domain adaptation techniques.

Our focus on training corpus as a variable also provides interesting results. We require better methods for evaluating NER between different corpora, and have begun developing tools for inter-corpus analysis.

Finally, Wikipedia’s breadth may allow us to reappraise these methods to specific domains such as biomedical NER, or to languages other than English.


Acknowledgements
Many thanks to my inspired supervisors, who best go hand-in-hand; and to my fellow Language Technology students for their comments, company and chocolate.

References
Flexible manipulation of sequence numbers in an interactive phylogeny analysis environment

Author: Bo Pang, bpan2323@mail.usyd.edu.au
Supervisor: A/Prof Bing Bing Zhou
School of Information Technologies

1. Motivation
Sequence number manipulation occurs frequently in phylogenetic studies. However, most existing phylogeny tools handle a fixed set of sequences in each tree building process. Biologists have to manually deal with sequence adding or removal from the current group, and restart the entire tree building process. This is a tedious job not only for biologists, as well as for the phylogeny analysis tools. A large amount of computation resources are wasted in those repeated calculations.

This project addresses the above limitations. Sequence insertion operation and sequence deletion operations are effectively incorporated into a web-based interactive tree analysis platform, named Qphyl.

2. Contribution
- Flexible sequence manipulation operations are integrated into the Qphyl system, including sequence adding, sequence deletion and their various combinations
- The interactive property is retained for the inferred phylogeny with an altered sequence number
- Computation cost is significantly reduced when analyzing diverse sub groups from homologous sequences
- An new insight is opened for several sequence manipulation oriented applications, such as rooting phylogeny.

3. Architecture
The system has three major components: front-end web-based interface, service nodes and backend execution engines. Front-end web-based interface is where users make sequence alternation request to the system, and get the result tree represented as a tree diagram. Ajax technology is applied to facilitate the communications between client and server, as well as the manipulation of DOM. Service nodes provide access interfaces to clients, and the additional functionalities include session management, workflow management, job scheduler, and job management. Execution engine execute the tree building requests coming from the service nodes. A quartet-based algorithm[1] is in support for phylogeny reconstruction.

4. Sequence adding operation
Sequence adding operation is achieved through five major activities: new sequences uploading, additional quartet weights generation, global quartet weight matrix updating, new tree reconstruction, and new tree rendering. Computation cost is saved from additional quartet weights generation. Only the ones relate to the newly added sequences are generated, instead of a complete set.

5. Sequence deletion operation
In sequence deletion operation, the involved tasks are: grabbing of the ID numbers of sequences to delete from web-browser, sequence data updating, ID correction of quartet weights, global quartet weight matrix updating and new tree rendering. In comparison with sequence adding operation, sequence deletion operation saves the time for generating additional quartet weights. Previous data records are applied after ID modification.

6. Experiments and results
Experiments are conducted from three aspects: Accuracy testing, performance analysis, and example application. Benchmark data set from LIRMM is used.

6.1. Accuracy testing
In comparison with a n-taxa tree from the original system, it is testified that an equivalent n-taxa tree is derived from a (n-m)-taxa tree through sequence adding operation, or from a (n+m)-taxa tree through sequence deletion operation.

6.2. Performance analysis
Sequence manipulation functions significantly improve the performance of the original system in dealing with sequence number changes. Sequence adding operation outperforms the original system when reconstruct a n-taxon tree after previous (n-1)-taxa tree [Figure 1]. The advantage becomes significant as n increases. Moreover, with regard to building a n-taxon tree from a (n-m)-taxa tree, for example n=40 [Figure 2]. The advantage of sequence adding operation over the original system reaches the peak when m=1, and decreases as m grows bigger.

Sequence deletion operation costs almost nothing in building a n-taxon tree from (n+m)-taxa tree regardless of what m value is. (e.g. 10 sec for 60-tree to 40-tree)

6.3. Example application
Tree rooting serves as an example for the applications to the sequence manipulation operations. Flexible sequence adding and deletion operations provides biologists an effective way in finding the root for homologous sequences.

7. Conclusion
Sequence adding operation and sequence deletion operation are successfully integrated into the Qphyl system. These new feature enable biologists to manipulate sequence number of the current group under analysis flexibly and effectively.

8. Future work
- Multiple sequence format support-chain Readseq into the system to transform numerous sequence data formats into PHYLIP before calculation starts
- Multiple sequence alignment support-complete the entire phylogeny inferring process effectively

9. References
Introduction
Database replication is an important technique to achieve high levels of performance, reliability and availability by providing several database replicas on multiple machines. There are several replication protocols available which follow a number of different approaches. Some approaches offer good performance with little guarantees given to availability or reliability while others provide reliability and availability at the cost of relatively poor performance.

Approaches are typically defined by two parameters [1], the first specifying how updates are coordinated and is either Eager (synchronous), or Lazy (asynchronous). The second specifies to whom clients can send updates to and is either Update Anywhere (multi-master) or Primary-copy (Master-Slave).

A recent idea [2] has been to combine concurrency control (in the form of snapshot isolation) with replication control in order to maximise performance and provide reliability by offering 1-copy generalised snapshot isolation. This approach has been successfully applied to both Eager and Lazy protocols. This research focuses extending a Lazy SI Update Anywhere approach with mechanisms to cope with failures.

Lazy Update Anywhere Replication
Because of the asynchronous nature of the protocol:
- Reliability Suffers
  Updates may be lost if we switch masters.
- Availability Suffers
  Wait until master comes back online.

Current approaches:
- Offer great performance
- Do not deal with failures

Research Objectives
Increase the robustness of Lazy Snapshot Replication by providing recovery in the event that the master node fails & mechanisms to dynamically extend the cluster and to cope with a complete failure:
- Without huge sacrifices to performance
- While maintaining Reliability

And evaluate these techniques in terms of:
- Their impact on performance
- Redundant data that must be kept around
- Comparison to other protocols

Architecture
- Lazy Update Anywhere Approach
  High performance; clients may send updates to any node.
- No Middleware
  The system does not suffer from the costs of a layer in between the clients and the DBMS.

The architecture exhibits all desirable characteristics
- Performance – Asynchronous updates
- Reliability - 1-copy generalised SI
- Availability – Able to tolerate the failure of the master node, or any number of replica nodes.

Contribution 1: Online Recovery
- The system operates with a single master node. This presents a single point of failure.
- There must be a mechanism to replace this single master in the event it fails. This mechanism must also ensure the reliability of the system.
- During normal operation, some redundant data must be maintained.
- Three adjustments are needed before the system can resume normally:
  1. The node containing the latest update from the failed master assumes the role of master and begins coordinating recovery.
  2. The new master rebuilds the failed masters propagation buffer up to the point of the last acknowledged update by coordinating with the remaining nodes. It then applies updates from the failed masters propagation buffer (highlighted in gray) and propagates the updates.
  3. The new master installs a new master id scheme.

Contribution 2: Offline Recovery
- Master can rebuild post-failure then distribute copies of database (slower).
- Coordinated Checkpointing & Recovery (quicker).

Contribution 3: Dynamic Scalability
- A new node must be able to join the cluster without interrupting operation.
- Install a snapshot while it buffers updates. Then apply updates until caught up.

Implementation
A prototype has been implemented with the PostgreSQL open source database engine, utilising the open source Spread library to provide reliable multicast and group membership notifications. PostgreSQL was chosen because it is an open source project, allowing modification inside the database engine and it provides a multi-version storage management that keeps updated records around which is a requirement of the protocol.

Scalability Characteristics
- System must maintain additional data
- How does this affect performance & scalability?

Conclusion
The Lazy Snapshot Replication algorithm was modified to keep additional data around to cope with several types of failures. These new fail safety algorithms were designed to minimise the impact on system performance while providing k-safety from slaves and protecting from master failure. Initial results are promising that online recovery of a failed master is under a second.

Acknowledgements
- Supervisor: Dr Uwe Röhm
- Michael Cahill
- SIT Database Research Group

References
Introduction & Goals
Existing operating systems fail to support users in identifying important relationships between their own documents. Consider the following scenario.

John is working on a conference report. Initially, he stores his notes in a file named Notes.tex. Later, he creates a new file, Final Report.tex, based on his notes. At this point, he may reasonably associate the two documents within a mental model of his personal filesystem.

Modern operating systems will treat these two documents as unrelated. However, such relationships can be important for finding the right file for a particular task, or when an activity calls for key related files that were used to create an existing file.

Days later, John is working on Final Report.txt. He realises that an important figure is missing from his report – he probably deleted it accidentally. However, he can’t remember exactly which notes he originally used to create this document, as they have already been filed away into another directory which contains presentation notes from all conferences held in 2008.

This poster describes Family, an architecture which aims to support users in seeing the temporal sequence of files that are part of the history, or lineage, of any nominated file. We also identify key conceptual models that constitute the cases of important family relationships between nodes. These correspond to high-level conceptual user models that constitute the cases of important family relationships between notes (as they represent an immutable state). We have identified four key models, listed below.

1. Document copy or ‘save as’
   This model maps to a simple user action of copying an existing file, or using ‘save as’ on an old file, to create a brand new file with the contents of the existing file.
   This is very useful for simple backtracking.

2. Document changes (saves) over time
   As a user saves his or her work over time, each state of a file can be uniquely identified, important for identifying revisions and other relationships to specific historical states.

3. Related work, copy and paste between documents
   By identifying documents both in use at similar times, as well as those with content swap, users can more effectively backtrack through the family of their documents.

4. Files sent to, and returned from, outside the machine
   The architecture of Family extends to creating relationships with documents moved to multiple machines or multiple users. However, this model and its relationships are semi-formal, as its state cannot be properly observed in this environment.

At t0, Zoe creates the file, Notes. She then edits the file, saving a revised version at t1. Zoe then creates the file, Report, by copying Notes using a command-line utility at t2. At t3, she updates her original file, but then copies parts of its content to her two new files, Report and Abstract, before saving them both. Her colleague is interested in annotating her final abstract, so at t4, she sends it to email. At t5, the file is returned, but not after undergoing an undefined set of changes.

Zoe’s example is illustrated in Figure 3.

Approach
The term document represents a user view of an idea or concept over time, regardless of its actual implementation within an operating system. A file, therefore, is the logical progression of this concept into the technical storage on a file system. Finally, a node represents each possible immutable state of this file — as a single file may be changed and saved by end-users many times throughout its lifetime. As such, a file may map to many nodes, but only the latest node — or current file state — is currently ‘active’.

John’s document, Final Report.tex, contains the markup for a final conference report. The document is backed by a single file created on the 15th of October, and which still exists now. He has saved his document many times as he makes changes, so it has existed in several immutable states, but there is only one version of the file he can now access.

This analogy is illustrated in Figure 2.

Conceptual Models
The broad aim of Family is to identify key conceptual models that constitute important family relationships between nodes (as they represent an immutable state). We have identified four key models, listed below.

1. Document copy or ‘save as’
   This model maps to a simple user action of copying an existing file, or using ‘save as’ on an old file, to create a brand new file with the contents of the existing file.
   This is very useful for simple backtracking.

2. Document changes (saves) over time
   As a user saves his or her work over time, each state of a file can be uniquely identified, important for identifying revisions and other relationships to specific historical states.

3. Related work, copy and paste between documents
   By identifying documents both in use at similar times, as well as those with content swap, users can more effectively backtrack through the family of their documents.

4. Files sent to, and returned from, outside the machine
   The architecture of Family extends to creating relationships with documents moved to multiple machines or multiple users. However, this model and its relationships are semi-formal, as its state cannot be properly observed in this environment.

Evaluation
A set of activity sequences was designed to test how well Family can identify all family members associated with a file. These sequences were undertaken both with no background noise — that is, other unrelated activity being parsed at the same time — and an ‘unparsed’ fashion where Family was not sand boxed to specific test data. On several experiments, Family succeeded on the conceptual models 1, 2, and 4, but failed on case 3. The existence of background noise did not affect the results. While Family does technically support the relationships concept described in case 3, we still need to refine agents capable of providing relevant evidence.

Conclusion
Family provides a new way to help users locate files that may be perceived as related within their own mental model. These relationships are temporal, and will typically be based on a file’s ancestors or descendant. In a broader sense, this information refers to how files were created and how they relate to the creation of others. The relations that Family derives may also be used as a set of source data for higher-level applications. For instance, version control software could use it to identify changes to files implicitly. This could also be combined with Apple’s backup software, Time Machine, in order to find related files which have been backed up in the past.

Acknowledgements
Thanks to Greg Darke and James Bunton for their testing and support on the low-level components of Family. Also, thanks to the Apple University Consortium (AUC) in Australia for supporting my thesis through both an Innovation Seeding Grant, involving Apple hardware and ongoing advice, along with a 2008 second-semester Honours scholarship.

Implementation
Family maintains information about the unique state of every file throughout its history, such as the date and time of last modified date. It also ‘tags’ each file in order to track its movement and changes throughout local or remote file systems. Internally, Family is primarily implemented in Python; but it also relies heavily on features in Mac OS X and a custom kernel extension, implemented in C, that provides a wide array of raw data to user-space.
1. Aim of the Project
To create a multi-resolution data structure for line models, in order to render them in real time, using techniques such as hierarchical view culling and adaptive simplification.

2. Introduction
Many 3D models today are very complex, and it is prohibitive to render them at full detail in real-time (~30Hz). The models must be simplified before rendering.

For massive and complex models, there is a huge disparity between foreground and background detail levels. An adaptive simplification scheme can retain detail in the foreground and simplify aggressively in the background. This is especially true for terrain rendering.

3. Line Models
Line models are a class of models that are composed of thin curves that exist in 3 dimensions. Examples of this include hair, fur, grass and foliage. These models are becoming more popular in recent years, as renderers become powerful enough to deal with these complex forms.

Even so, simplification is still often required. In real-time rendering, the following properties are desirable:
- Multi-resolution model, a collection of models of varying levels of simplification
- Hierarchical data structure, to perform adaptive view-dependent simplification and culling
- Simplify to minimise perceived error

Line models can exhibit strong inter-line similarities and clustering behaviours. It is naive to perform curve simplification on the lines independently, without considering inter-line similarity.

For this reason, a segment-centric approach to simplification is proposed. Segment-centric simplification considers similarity between the piecewise linear segments of lines, instead of the lines themselves.

4. System and Approach
The system is composed of two parts: an offline data pre-processor, and a real-time renderer.

The pre-processor takes as input a line model, and builds a multi-resolution hierarchy. The data structure is an adaptive tree hierarchy of bounding spheres. At every node, the line segments represented by its children are merged into an averaged representation in order to minimise perceived error.

Line segments have many similarity metrics, such as proximity, orientation, and spatial relationship. This means simple partitional hierarchies are inadequate, and an adaptive approach needs to be used.

K-nearest-neighbour clustering was used to construct the hierarchy. This allows good mergers whilst keeping the time and space complexities low, allowing processing on datasets of millions of primitives.

5. Results
The most appropriate measure of the effectiveness of simplification is to compare the output images. For real-life line models such as hair, where there are large uniform areas and the model is not overly complex, the model can be rendered in less than 10% of the time and with less than 5% in image space error. Errors are concentrated on the silhouette.

The error is much worse for synthetic models of high complexity, with up to 20% image space error for similar reductions in render time. Errors are evenly spaced throughout.

6. Discussion
Although error is low for real-life models, the tendency for errors to manifest on noticeable areas, especially the silhouette, is visually jarring and qualitatively bad. A weighting metric that preserves detail in important parts can significantly improve visual quality.

Geometric errors suffer because of the simplistic segment-centric pre-processing. On the other hand, shading characteristics is preserved throughout simplification, greatly improving visual quality.

7. Future Work
- This work used a simplified shading model. In reality, dense line models produce volumetric effects such as self-shadowing. A more realistic shading model can enable better analysis of simplification error.
- In practice, many line models have characteristics that allow tailored simplification – e.g. hair models have very similar strands. A possible improvement is to detect these characteristics through careful analysis to create better simplifications.

Model of blonde hair, original vs. simplified, 300x300 image size. Left: 1070k segments, 719ms; right: 32k segments, 32ms, 8% image-space error.

"Comet" model, original vs. simplified, 1400x1050 image size. Left: 3m segments, 1812ms; right: 1.5m segments, 1015ms, 14% image-space error.
Background
In the current era of information technologies, database is everywhere. Almost every large corporation requires a database system to maintain their business operations. The dependency on the performance of a database varies and would be rather small if a business uses one only to keep its histories. However, for businesses with their profits predominantly created from the processing of online transactions, the performance becomes a very crucial issue. There have always been numerous attempts to improve the performance of a database system as well as to invent a set of defined procedures to measure the performance correctly and fairly. This system of evaluating the performance is called a ‘benchmark’.

Commercial products use their own benchmark or so-called a ‘standard benchmark’, which is created by an organization of people participated from various database hardware and software vendors, to illustrate its superior performance over other products. From an academic point of view, researchers frequently use a benchmark to reinforce the significance of their work. Unfortunately, there is no single universal benchmark that suits for all purposes. Once Gray [3] identified that each benchmark should be domain-specific along with other characteristics that a benchmark must have. Therefore, standard benchmarks have been created to provide a domain-specific benchmark, yet that still can be generally accepted and used by a wide range of systems. Still, this resolves the issue only partially and not enough to serve individual needs for own benchmarks.

It is a tedious task to design a database benchmark. A database benchmark should be accurate with results and fair to target systems so that it does not favour a system of a specific kind. The designer also needs to be aware of the repeatability of the benchmark. It would be the most likely case that the benchmark gets run repeatedly during the evaluation process and the result must be reproducible by others who want to confirm it.

Objectives
Database Performance Evaluation Framework (DBPEF) is an approach to reduce and simplify the requirements necessary for building and running a database benchmark. The objectives of this research are:
- To identify core aspects that most database benchmarks share.
- To design a framework for database benchmarks reducing efforts required to design and supporting their implementation.

The significance of this research primarily lies with the framework’s ability to capture the core aspects of database benchmarks to assist them, not the framework itself.

Database Performance Evaluation
The implementation of a database benchmark differs with the context of a target system-under-test (SUT). However, the measurements used to evaluate the performance of a system are usually same. They are:
- Throughput: Number of successful request/transactions processed by SUT in a time unit. The measurement is typically done in tps (transactions per second) or tpm (transactions per minute).
- Response time: Time duration SUT takes to answer a request or complete a transaction. This is typically measured in milliseconds.
- Surely there are special measurements that a benchmark uses, but mostly they are an extension of the two main measurements.

Most benchmarks have similar procedures in common. This is the result of a continuous effort in the database community to describe the essential requirements. [1] [2] The main focus of this research is an identification of non benchmark-specific aspects.

Database Benchmark Processes
There are several steps that a database benchmark takes in its evaluation process:
1. Set-up: initialises the SUT by creating appropriate tables and generating data into tables for a benchmark.
2. Run: executes predefined transactions with random parameters. Metrics are measured here.
3. Tear-down: cleans the SUT up by removing generated data and tables and terminates connections with the SUT.

Database Benchmark Details
Major benchmark-specific aspects are generally designed in the following order:

Result
The framework captures following aspects of a benchmark:
- Emulation of clients (i.e. workload generation) done by DBPEF locally using multi-threads or in a distributed computing environment where multiple client machines can involve in a benchmark process.
- Automation of benchmark processes as long as the implementation details are correctly provided.
- Measurement of throughput and response time.
- Determination of steady-state.
- Division of parameters that need to be altered frequently by a tester from the benchmark. (e.g. MPL, duration of benchmark, etc)

Conclusion and Further work
We have shown that there are common aspects in database benchmarks and it is possible to encapsulate them as a framework. The framework reduces the effort required to create and implement a benchmark significantly. Further work involves a detailed discussion on the resolution of timing functions depending on platforms and even programming languages, which can have some serious effects on the accuracy of benchmark results. We also would like to study further a methodology of determining a steady-state during the benchmark process so that the framework can intelligently detect a steady state and start the measurement regardless of the type of a benchmark it is processing.

Acknowledgements
I would like to thank my supervisor, Dr. Uwe Roehm, for his support and guidance throughout this project.

References