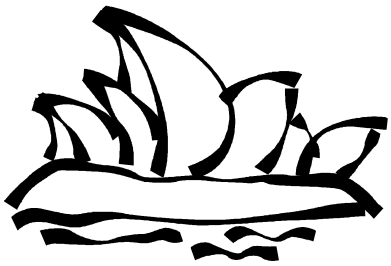
# ADCS 2005

# Proceedings of the Tenth Australasian Document Computing Symposium,

12 December 2005

Edited by
Judy Kay, Andrew Turpin and Ross Wilkinson



Proceedings of the Tenth Australasian Document Computing Symposium,
held at The University of Sydney
12 December 2005.

Published by the
School of Information Technologies,
The University of Sydney,
NSW 2006, Australia.

Editors:

Judy Kay
Andrew Turpin
Ross Wilkinson

ISBN 1-86487-787-1

http://www.cs.rmit.edu.au/~aht/adcs2005


# ADCS 2005
# The Tenth Australasian Document Computing Symposium

## Sydney, Australia
## 12 December 2005

### Chairs' Preface

These proceedings contain the papers of the Tenth Australian Document Computing Symposium hosted by, and held within, the School of Information Technologies at The University of Sydney.

The two keynote addresses, eleven papers and six abstracts reflect the breadth and interest of the Australian research community in the area of document computing.

The eleven papers here were selected from thirteen submissions. Every paper had three peer reviews. Submitted papers were anonymously reviewed at their full length by experts in the area. Dual submissions were explicitly prohibited.

The members of the program committee and the extra reviewers deserve special thanks for their contribution to ADCS2005. Reviewers not listed among the program committee include Halil Ali, Michael Cameron, Richard Cole, Jon Ducrou and Mingfang Wu. We would also like to thank the University of Sydney for their involvement in 2005 as sponsor and host of the event. The event is also supported by the ARC HCSNet.

The symposium includes many formal presentations, but perhaps its greatest benefit lies in the opportunity it provides for document computing practitioners to get together informally and to share ideas and enthusiasm.

Judy Kay
Andrew Turpin
Ross Wilkinson

## Symposium Chair

| | |
|---|---|
| Judy Kay | The University of Sydney |

## Program Co-chairs

| | |
|---|---|
| Andrew Turpin | RMIT University |
| Ross Wilkinson | CSIRO |

## Program Committee

| | |
|---|---|
| Vo Ahn | The University of Melbourne |
| Theresa Dirndorfer Anderson | University of Technology, Sydney |
| Peter Bruza | Queensland University of Technology |
| Bob Colomb | University of Queensland |
| Stijn Dekeyser | University of Southern Queensland |
| Peter Eklund | University of Wollongong |
| Tom Gedeon | Australian National University |
| David Hawking | CSIRO |
| Judy Kay | University of Sydney |
| Rob McArthur | CSIRO |
| Alistair Moffat | University of Melbourne |
| Gitesh K. Raikundalia | Victoria University |
| Falk Scholer | RMIT University |
| Saied Tahagohghi | RMIT University |
| Jamie Thom | RMIT University |
| Anne-Marie Vercoustre | INRIA, France |
| William Webber | The University of Melbourne |
| Justin Zobel | RMIT University |

## ADCS Advisory Committee

| | |
|---|---|
| Peter Bruza | Queensland University of Technology |
| David Hawking | CSIRO |
| Judy Kay | University of Sydney |
| Alistair Moffat | University of Melbourne |
| Ross Wilkinson | CSIRO |
| Justin Zobel | RMIT University |

# Contents

# From Non-Segmenting Language Processing to Web Language Engineering

Virach Sornlertlamvanich
Thai Computational Linguistics Laboratory (TCL), NICT, Thailand
virach@tcllab.org

It is interesting to look at the statistics of the online languages reported by the Global Reach (www.global-reacg.biz). In September 2004, it was reported that the top six online language populations were English 35.2%, Chinese 13.7%, Spanish 9.0%, Japanese 8.4%, German 6.9%, and French 4.2% while the web contents were English 68.4%, Japanese 5.9%, German 5.8%, Chinese 3.9%, French 3.0% ,and Spanish 2.4%. There are some changes in ranking between the online language populations and the existing of the web contents. However, English is still the majority language used in the online community. Many efforts have been making to prevent the fall-off in using of other languages, especially the less computerized languages. It is said that there are about 7,000 languages using in all over the world. At the same time the less computerized languages are disappearing. The Rosetta Project (http://64.81.54.21:8080/live/) is a global collaboration to build an online archive of all documented human languages. The Language Observatory Project (www.language-observatory.org) initiated by Nagaoka University of Technology to search for the disappearing languages.

To deal with languages as many as we can find online, it is much more efficient to consider the language independent approaches. The big difference between segmenting languages (i.e. English and other European languages) and non-segmenting languages (i.e. Thai, Lao, Khmer, Japanese, Chinese and a lot of Asian languages) in the existing of word boundary marker causes the change in language processing. Most of the current approaches are based on the assumption that words are already identified disregarding the existing of the word boundary markers. The research on word boundary is separately conducted under the topic of word segmentation. On contrary, we proposed some algorithms to handle the non-segmenting languages (Virach 2005a, Virach 2005b) to establish a language independent approach.

In our recent research, we proposed a language interpretation model to deal with an input text as a byte sequence rather than a sequence of words. It is an approach to unify the language processing model to cope with the ambiguities in word determination problem. The approach takes an input text in the early stage of language processing when the exhaustive recognition of total word identity is not necessary. In our research, we present the achievements in language identification, indexing for full text retrieval, and word candidate extraction based on the unified input byte sequence. Our experiments show comparable results with the existing word-based approaches.

In our statistical-based word extraction research (Virach et al. 2000), it was reported to yield about 30% of the total word candidates being the unregistered words of a published dictionary, when the recall threshold was set to 56%. Character-based mutual information and entropy provided significant information to C4.5 algorithm for selecting appropriate candidates for words. The approach greatly supported the process of developing a dictionary, and later was extended to fulfill a dictionary-less search engine (Virach et al. 2003). The search engine had introduced a word score as a heuristic value to determine the word likelihood of a string. The word score was a normalized value of a mutual information value. The minimum score of the left and right hand side of a string in question was assigned as the word score of the string. Based on the proposed approach, we successfully implemented a multi-lingual search engine with minimum modification.

Language identification (Canasai et al. 2005) is yet another challenging task when it is done without any parsing knowledge. Byte sequence is the only magic key in our approach to determine the language of the input text. We introduced string kernel for this language identification task. We conducted experiments using 2 kernel classifiers i.e. centroid-based and support vector machine (SVM) methods. The accuracy of identification was acceptable for both methods. The accuracies reached 95 percent with only 10 training sets (2 KB per set). It was also found that the simple centroid-based classifier is comparable to the SVM classifier based on the string kernel.

Our approaches had been proven effective under the Thai language and the multi-lingual environment of 16 European and 4 Asian languages including Thai, Chinese, Japanese, Korean, English and many other

European languages. We are expanding our corpus for conducting our experiments under the environment of a large number of languages.

Based on the successful results of word extraction, language identification and language independent indexing for search engine, we are conducting an experiment of the collaborative crawler on the high speed link (45 mbps) between Thailand and Japan. This collaborative work will provide an infrastructure for collecting web contents to study about the web language. The language together with its encoding of every webpage will be automatically identified and indexed to make the archive. Collaborative search engine will then go through all archives in all registered sites to present the ranked search results for any particular requests in any languages. The reports on the web languages from any perspectives can also be constructed by the proposed web language engineering.

**Reference:**

Virach Sornlertlamvanich. Implementations that Unify the Language Processing, Proceedings of the 9th NCSEC, University of Thai Chamber of Commerce, Bangkok, Thailand, pp. 1053-1062, 27-28 October, 2005.

Virach Sornlertlamvanich. Statistical-Based Approaches for Non-Segmenting Languages, Proceedings of Pacific Association for Computational Linguistics (PACLING), Meisei University, Tokyo, Japan, pp. 75-84, 24-27 August 2005.

Virach Sornlertlamvanich, Tanapong Potipiti and Thatsanee Charoenporn. Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm. Proceedings of the 18th International Conference on Computational Linguistics (COLING2000).

Virach Sornlertlamvanich, Pongtai Tarsaku, Prapass Srichaivattana, Thatsanee Charoenporn and Hitoshi Isahara. Dictionary-less Search Engine for the Collaborative Database, Proceedings of The Third International Symposium on Communications and Information Technologies (ISCIT-2003), Songkhla, Thailand, 3-5 September 2003.

Canasai Kruengkrai, Prapass Srichaivattana, Virach Sornlertlamvanich, and Hitoshi Isahara. Language Identification Based on String Kernels, Proceedings of the 5th International Symposium on Communications and Information Technologies (ISCIT-2005), Beijing, China, October 12-14, 2005.

**Document Processing Using Formal Concept Analysis**

*Peter Eklund*
School of Economics and Information Systems
The University of Wollongong
Australia
*peklund@uow.edu.au*

Formal concept analysis (or concept lattices) have repeatedly been applied to document processing and clustering over 15 years. But what are the outcomes? In this talk I will trace the history of the idea and show some recent results. The main emphasis of the talk will be on computability, interface design and usability issues. Some of the most exciting recent results demonstrate the idea by applying it to image browsing. In the image case sub-context clustering using concept lattices is based on various combinations of meta-data tags, colour and shape attributes using MPEG-7. These outcomes give arguably the best platform for concept lattices to date.

# Cross Training and Under Sampling in Categorization of Company Announcements

*Cheng G. Weng*
School of Information Technologies
The University of Sydney
Sydney NSW 2006, Australia
*cheng@it.usyd.edu.au*

*Josiah Poon*
School of Information Technologies
The University of Sydney
Sydney NSW 2006, Australia
*josiah@it.usyd.edu.au*

**Abstract** *To process the documents in a share market is crucial. It is because financial activities are socio-economic driven and text documents contain a lot of valuable information. In this paper, we focus on one of these documents, the Company Announcement. Each of these documents requires to be labelled as price sensitive or not before presenting to the general public. In our experiments, we study two specific issues in this text categorization, namely the effectiveness of a feature vector obtained from the corpus belonging to another market sector and the imbalanced nature of the dataset. Our results indicate that the classification can benefit from a different (but related) set of corpus because of a more diversified and generalised nature of the feature set. Regarding the skewness of the dataset, the under-sampling of the majority class in the training process does not have a strong effect on the performance in the test set, while keeping the computational cost minimised.*

**Keywords** Document Management, Text Categorization

## 1. Introduction

Information overload is a major problem in this new age of World Wide Web, and the majority of these information are in textual form. The need to manage and utilize textual information has led to the birth of text mining research. In the financial domain, many important decisions making are based on the assessment of text documents, and the timing of these decisions is also crucial. It has, therefore, attracted researchers to work on this area [4, 6, 8].

In this paper, we used company announcements obtained from the Australia Stock Exchange (ASX) website and tried to explore more effective ways of training on text documents. The next section describes some related work, follow by a section that introduces our dataset, then section 4 elaborates the motivations for the experiments describes in section 5. We present the experimental results in section 6 and discuss the observations in section 7. Finally, we will conclude with some future work.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

| Dataset | Sen | Nonsen | Total | Sen% |
|---|---|---|---|---|
| Website | 1472 | 9604 | 11076 | 13% |
| Signal G | 46530 | 90100 | 136630 | 34% |

Table 1. Dataset comparison with [1]. *Website* is the dataset we used for our experiment, and *Signal G* is the name of the dataset used in [1].

## 2. Related Work

Our approach is based on the traditional statistical text categorization process, which transform text documents into word vectors. Essentially, treating the documents as 'bag-of-words' and ignore other linguistic information. It is a rather superficial approach but it has been shown to be effective in practice [7].

To the best of our knowledge, there was one closely related work done by Calvo and Williams [1]. They also used the announcements, but they had access to a larger dataset, because of their affiliation with the Capital Market CRC[1], which we did not have. They compared the performance of different machine learning algorithms, and concluded "the good performance shows the possibility for commercialization". Table 1 shows the differences of our dataset to theirs.

Since our dataset has a skewed class distribution, we used the same experimental setting as in [1] to check the representativeness of our company announcement sample. We used random under-sampling to accommodate the class imbalance of our dataset, and this sampling technique has been shown useful for non-textual datasets with class imbalance [2].

## 3. Dataset

Before describing our motivations for the experiments, we first provide some background knowledge about our dataset, the ASX company announcements dataset. These company announcements are manually categorized by their market sensitivity, which is either

[1] It is a joint research centre with stakeholders coming from the industry and the academic institutions, of which ASX is one of the industrial partners.

| Code | Sen | Nonsen | Sen% | Unique | Sector |
|---|---|---|---|---|---|
| ANN | 21 | 273 | 7% | 11703 | H |
| CSL | 25 | 172 | 13% | 12118 | H |
| SIG | 13 | 86 | 13% | 7229 | H |
| VCR | 30 | 132 | 19% | 6232 | H |
| ANZ | 54 | 637 | 8% | 27312 | F |
| CBA | 38 | 1816 | 2% | 25035 | F |
| MBL | 97 | 2523 | 4% | 68943 | F |
| NAB | 80 | 674 | 11% | 24219 | F |
| AWE | 137 | 428 | 24% | 11374 | E |
| ROC | 164 | 460 | 26% | 12496 | E |
| STO | 265 | 455 | 37% | 11924 | E |
| WPL | 113 | 310 | 27% | 12281 | E |
| BHP | 148 | 325 | 31% | 22168 | M |
| BSL | 39 | 329 | 11% | 16395 | M |
| ORI | 65 | 350 | 16% | 10495 | M |
| RIO | 70 | 158 | 31% | 17692 | M |
| ERG | 46 | 95 | 33% | 14998 | I |
| IFM | 16 | 118 | 12% | 9966 | I |
| MYO | 23 | 189 | 11% | 882 | I |
| VSL | 28 | 74 | 27% | 6460 | I |

Table 2. Company Statistics.

"market sensitive" or "market non-sensitive". The market sensitivity of an announcement depends on its predicted effect after release to the general public, and this will be judged by the experts in the ASX. If the information could potentially have significant impact on the market, they will be labeled as market sensitive, otherwise market non-sensitive. Because of the subjective nature and the difficulty of knowing the exact impact of an announcement, the experts are more conservative when labeling a document as market sensitive. Therefore, while some announcements are labelled as market sensitive, it is not a guarantee that it did have a significant impact on the market. A sensitive document is considered as a rare and important event in this task, so the cost associated with misclassified sensitive document is high, e.g. ill-informed investors may miss out their best buy/sell timing for their stocks.

The announcements are publicly available in PDF format from the ASX website[2]. We have collected announcements for 20 listed companies on ASX200, each with more than 2-year worth of data, from early 2003 to early 2005. We kept the corpus separated by individual company, and the size of the whole corpus is about 175 MB. Each company belongs to a market sector, which is defined by the Global Industry Classification Standard (GICS).

Table 2 shows the basic statistics for each company. The 'Code' is the trading code of the company on ASX, 'Sen' is the number of sensitive documents, 'Nonsen' is the number of non-sensitive documents, 'Sen%' is the percentage of sensitive documents in the company, and 'Unique' is the number of unique tokens after pre-processing (described in section 5). 'Sector' is the market sector, where *H* stands for Health Care sector, *F* stands for Financial sector, *E* stands for Energy sector, *M* stands for Material sector, and *I* stands for Information Technology sector.

[2] ASX `http://www.asx.com.au`

## 4. Motivations

The general assumption is that one needs to use a set of training data that resemble the future test data, in order to obtain the best generative model that minimize the test error. We have, therefore, proposed to test this assumption in a text categorization task.

The text categorization task can be viewed as two steps: step one is the selection of features, which we call *TCFeature* (TCF), and step two is the modeling of the data after they are being represented in vector form using the features selected from step one, and we call this step *TCModel* (TCM). In the traditional approach, both TCF and TCM are performed on the same company. So our hypothesis is that a similar performance can be acheived when the TCF is operated on a different company but belonging to the same market sector as the testing company. From here on, the notation TCF and TCM will be used in the rest of the paper.

Next, we tested the effect of under-sampling by randomly remove the non-sensitive documents, the majority class, from the data. We assumed this would give us a better performance on the minority class, which is what we prefer.

## 5. Experiments

Although both our dataset and [1] came from the Australian Stock Exchange, they are different. Therefore, the aim of the first experiment (**Exp1**) is to find out if the difference will constitute any significant performance variance. For **Exp1**, we used the same setup as in [1], which was a typical process in text categorization. Starting with stopwords removal, the apply the Porter word stemming algorithm [5], perform the feature selection, index the documents with TFIDF [7], and finally build the model with a machine learner. They used document frequency for feature selection to select 1000 features, and Support Vector Machines (SVM) was one of the machine learner they have used.

For our other experiments, we used similar setting, except, we chose information gain as the feature selection method to select 3000 features and SVM as the machine learning method. These choices were made because our focus was not on the feature selection nor the machine learner, and also the previous studies have shown good results when applying the two methods in text categorization [7]. All experiments are evaluated with 10 fold cross-validation.

The second experiment was set up to compare the performance, when different companies are used for TCF. There are two parts to this experiment: in the first

| | Micro | | | Macro | | |
|---|---|---|---|---|---|---|
| | $p$ | $r$ | $F_1$ | $p$ | $r$ | $F_1$ |
| L | 0.82 | 0.82 | 0.82 | 0.80 | 0.79 | 0.80 |
| S | 0.90 | 0.90 | 0.90 | 0.79 | 0.72 | 0.75 |

Table 3. *Exp1:* Comparison with [1]. *L* is the results from [1], and *S* is our results.

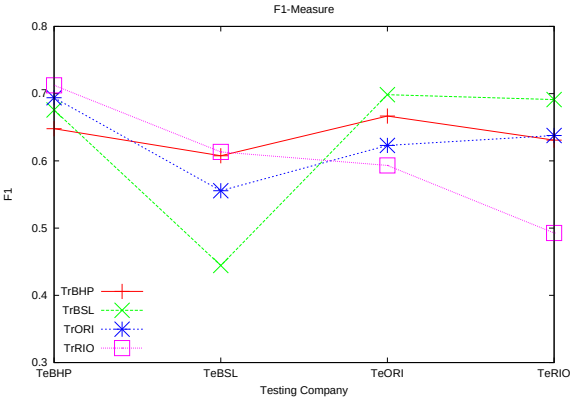



Figure 1. *Exp2a*: results for the Material sector. Each line represents TCF on a company, and x-axis is which company is used for testing.

part the different companies belongs to the same market sector as the test company (**Exp2a**). In the second part the different companies belongs to different market sectors as the test company (**Exp2b**). The control for this experiment is the one when TCF and TCM are both performed on the same company.

The third experiment tests the effect of random under-sampling of non-sensitive documents from a company. The setup is identical to the second experiment, except the data of each company will be reduced at the training stage. We tried 11 different reduction rates ranging from 0%, the original corpus, to 100%, where only sensitive documents are kept. Again, there are two parts to the experiment: the first part is when TCF and TCM are done on the same company (**Exp3a**), and the second part is when TCF and TCM are done on different companies (**Exp3b**). In the second part of the experiment, because TCF is independent of TCM, the features will remain consistent, while the training data for TCM gets the reduction.

## 6. Results and Evaluation

The micro and macro averages [7] was reported in [1], hence, we did the same calculation for comparison in **Exp1**. Table 3 shows we have a similar result, which suggests that our dataset is a compatible sample.

Due to space limitation, we will only report subset of the results for our experiments **Exp2** and **Exp3**. But this subset of the results is representative and the reported observations hold for all other results.

The chart shown in Figure 1 is one of the results for experiment **Exp2a**, while Figure 2 shows the results for **Exp2b**. We observed three phenomena:

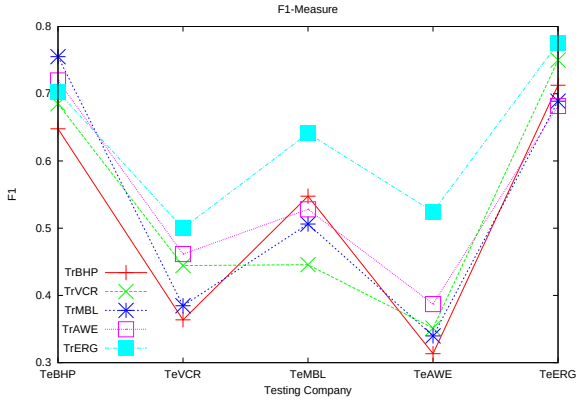



Figure 2. *Exp2b*: Results for TCF on companies in different market sectors. All 5 companies shown here belongs to different market sectors.

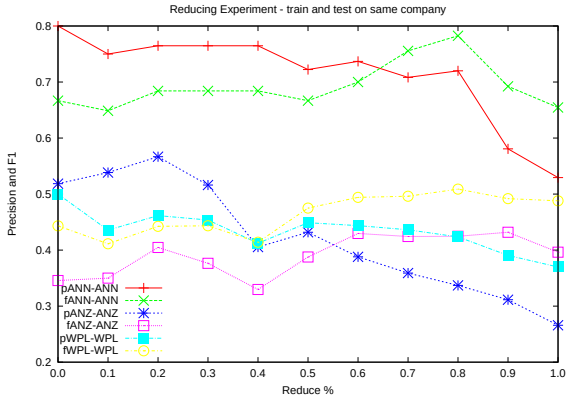



Figure 3. *Exp3a*: Reducing non-sensitive documents for ANN, ANZ, and WPL. *pANN-ANN* stands for the precision of TCF on ANN and testing on ANN, and *fANN-ANN* is the F1 measure. The same notation applies to others. The "Reduce %" is the percentage of non-sensitive documents removed.

**Observation** 1: Perform TCF and TCM on the same company does not always give the best results, which seems to be counter-intuitive. For all 20 companies, we found only 7 cases, where performing TCF and TCM on itself give above average performance. However, it was never the best, only 2 out of the 7 it was in top 5.

**Observation** 2: From Figure 2, the classification performance does not have significant difference whether the TCF was done on companies belonging to the same or different market sector.

**Observation** 3: The performance rises for certain companies but dips for another. The performance of a company being tested seems to be bound by some company dependent variable.

The third experiment **Exp3a** is shown in Figure 3, and Figure 4 shows the results for **Exp3b**. They have exhibit another 2 phenomena:

**Observation** 4: Reducing, up to 50% of, the non-sensitive documents does not deteriorate the performance. When the reduction is greater than

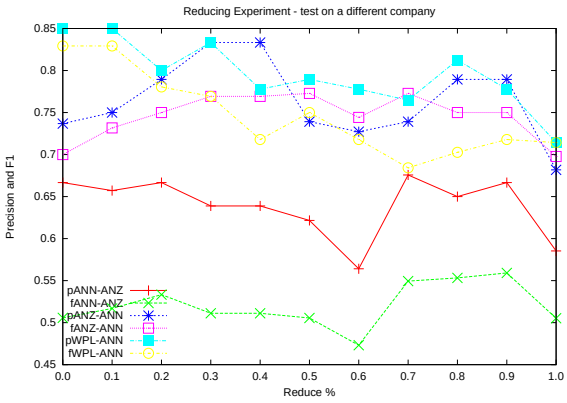



Figure 4. *Exp3b*: Reducing experiment for testing on other companies. The notation used here is the same as in figure 3.

50% the F1 measure does not drop, because the recall boost compliments the precision drop.

**Observation** 5: When the TCF is done on other companies, even if the non-sensitive documents are absent, the TCM can still be done with reasonable performance.

## 7. Discussion

In the previous section, we have shown a similar result compare with [1]. The higher micro average is due to the difference in performance of our two classes. The performance for non-sensitive class is much higher than the sensitive class. Next, we will discuss the observations made in our experiment.

Observation 1 suggests the feature vector can be constructed from another company has better generalization, i.e. the features are more diversified and generalised, and observation 2 suggests there is generic attribute that does not change across market sectors.

Observation 3 suggests the possibility of company dependent variables effecting the performance. So we attempted to look for a correlation between the performance and varies company dependent statistics, but none has shown a strong correlation for conclusion. The statistics we have tried are the class distribution, the number of documents, the percentage of sensitive documents, the number of unique tokens, the number of unique tokens in sensitive and non-sensitive document separately, and the overlapping of unique tokens of sensitive and non-sensitive documents.

From observations 4 and 5, we see a discriminative feature vector can maintain the performance even when the documents in the majority class is removed. The word usage can still be modeled without compromise the performance. When we looked at the features selected from different reduction rate in **Exp3a**, we found high overlapping of features among the feature vectors. This suggests similar features can still be selected without most of the non-sensitive documents.

## 8. Conclusion and future work

Our hypothesis in Section 4 aimed to explore the building a feature vector from a different corpus, and also the effect of random under-sampling on the dataset. We discovered even though the feature vector was constructed from a different corpus, they still give good performance, and frequently outperform the feature vector generated by the training corpus itself. We also found that random under-sampling of majority class does not deteriorate the F1 measurement, even when the majority class is completely removed. These empirical evidences suggests the possibility of a more effective text categorization process, by obtaining feature vectors from a better source, and model the word usage with a smaller subset of documents.

An interesting future work should be testing the same procedure on other standard text categorization copra and see if our findings still occurs. Also, the imbalance nature of the dataset would be another interest direction for further investigation [3].

## References

[1] R.A. Calvo and K. Williams. Automatic categorization of announcements on the australian stock exchange. In *7th Australasian Document Computing Symposium*, Sydney, Australia, 2002.

[2] N. Chawla, K. Bowyer, L. Hall, and P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. In *International Conference on Knowledge Based Computer Systems*, 2000.

[3] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–450, 2002.

[4] Antonina Kloptchenko, Tomas Eklund, Jonas Karlsson, Barbro Back, Hannu Vanharanta, and Ari Visa. Combining data and text mining techniques for analysing financial reports. *Intelligent Systems in Accounting, Finance and Management*, 12:29–41, 2004.

[5] Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[6] G. Pui Cheong Fung, J. Xu Yu, and Wai Lam. Stock prediction: Integrating text mining approach using real-time news. In *Computational Intelligence for Financial Engineering, 2003. Proceedings. 2003 IEEE International Conference on*, pages 395–402, 2003. TY - CONF.

[7] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[8] Dongsong Zhang and Lina Zhou. Discovering golden nuggets: data mining in financial application. *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 34(4):513–522, 2004. TY - JOUR.

# Recommending Geocaches

*Andrew Trotman* *Timothy Jones* *Chris Handley*

Department of Computer Science
University of Otago
Dunedin, New Zealand

*{andrew, tljones, chandley}@cs.otago.ac.nz*

**Abstract** *Players downloading GPS coordinates from the internet, hiking to the given spot, and hunting for a hidden box – this is the new sport of geocaching. Today there are nearly 200,000 such boxes in over 200 countries. With so many to find, a recommender is needed, one that takes into account not only the boxes, but also the geospatial and temporal nature of the sport.*

*A database of geocaches in the South Island of New Zealand is made by trawling a prominent geocaching web site. This is then used to estimate the home-coordinates (geospatial playing centre) of players. Predictions are verified against a set of correct coordinates solicited from players.*

*Several geocache recommenders are discussed and compared. The precision, computed using mean of mean reciprocal rank (MMRR), of each is measured. The best method tried is a collaborative filter using intersection over mean to find similar players and a voting scheme to recommend geocaches. This method is proposed as a replacement for the currently used distance from home-coordinate; doing so will increase the precision of existing systems such as geocaching.com.*

**Keywords** Information Retrieval.

## 1. Introduction

When the US president Bill Clinton announced the descrambling of the GPS satellite navigation system on the 1st of May 2000, he unwittingly also invented a new outdoor individual sport today known as geocaching.

The selective availability scrambling was removed on the 2nd of May 2000 and the next day Dave Ulmer hid a bucket of miscellaneous items (including a log book) in a forest outside Portland, Oregon. He published the coordinates on USENET and within a day the bucket had been found [9]. Within a month there were similar geocaches hidden in not only other US states, but also in other countries (including Australia and New Zealand). Today there are 196,250 caches in 217 countries [3].

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.**


This new sport is like orienteering; however, unlike orienteering, it is an individual sport. Players download longitude/latitude coordinates from a website (such as geocaching.com), go to the given location and then search for a hidden box. On finding it they open the box, log the find in the log book, then put the box back. Later they return to the website and log their find electronically. The sport can be played any time of the day or night, by anyone with a GPS receiver – there is no setup, no cleanup, and there are no teams.

Each new player expands the sport by hiding geocaches in places they enjoy visiting. Some players use the sport to swap walking tracks – they might, for example, hide a geocache on the ridge of a mountain. Other players might prefer obscure locations in big cities. Lunch-time players hide them in easy to get to places that make a good location for a lunch break.

This user preference brings both diversity and confusion to the sport. When geocaching in a new city (perhaps on holiday) a player is faced with several hundred geocaches from which to choose the few they might enjoy finding.

In this investigation we ask the question – is it possible to build a recommender for geocaching? But first we ask – is it possible to determine from behaviour where (geographically) players are playing?

We build a list of geocaches in the South Island of New Zealand by trawling geocaching.com.au[1]. Gaussian filtering is shown to be effective in home-coordinate estimate. Voting by similar players is an effective recommender; similar players are those with a high ratio of finds in common.

## 2. Recommender Systems

In a traditional collaborative filter recommender system such as MovieLens [13] an object's rating is predicted using a statistical analysis. For a given user the nearest other users are computed (perhaps using a *k*-nearest neighbour algorithm) and from that a weighted average of those users' ratings is used to rate the object. In effect, the rating a user will give the object is estimated using a weighted average of the ratings that similar users already gave it.

In a supermarket recommender [10], recommendations are made based on objects the user has pur-

[1] geocaching.com forbids trawling.

chased. For example, users who buy cheese and grapes are likely to also need crackers and wine. The recommendations are made by mining the shopping lists of customers. The purchases of all customers are collected together and data mining techniques used to find objects that are usually purchased together.

The domain of the recommender must be taken into account when choosing algorithms. Using a collaborative filter in a supermarket might tell us that a user would like a given brand of milk, but milk is milk regardless of brand and the user already knows this.

Equally, telling a user that if they enjoyed the first movie in a trilogy they should watch the others is futile – they already know this.

Recommending geocaches is quite unlike recommending supermarket purchases or movies for many reasons:

Players rarely return to the same geocache twice. This is quite unlike a supermarket where the same people usually buy essentially the same objects each visit.

Players cannot rate geocaches so it is not possible to predict a "movie rating" as there is no notion of rating.

In a supermarket all the objects on the shelf are available for purchase. In a movie recommender like Amazon.com all movies are also available for purchase. But this property does not hold for geocaching – just because a database is aware of a given geocache it does not mean the player can get to it. With a movie recommender like MovieLens, some very obscure items may be recommended, but no longer available for purchase [11], however with geocaching these objects are not obscure, they are geographically separated from the player. Recommending an Australian geocache to a New Zealand player is of little value – they cannot get to it!

New geospatial collaborative filtering algorithms are needed for this sport – algorithms that take into account the player's habits and recommend only accessible geocaches. We are aware of no such pre-existing algorithms and focus on such algorithms (both content based and user based) in this investigation.

## 3. Home-Coordinates

At present, geocaching.com recommends based on distance from a player's registered home-coordinate. This coordinate is the location that the player gives as the centre of their geocaching activity. Players are believed to use either their true residential home location or their work location (although this is anecdotal). Knowing the player's current location is essential for recommending any geospatially dispersed objects – without it, it is not possible to recommend close objects. These home coordinates are protected by geocaching.com and are not on geocaching.com.au (and neither site is ours) so estimates are needed.

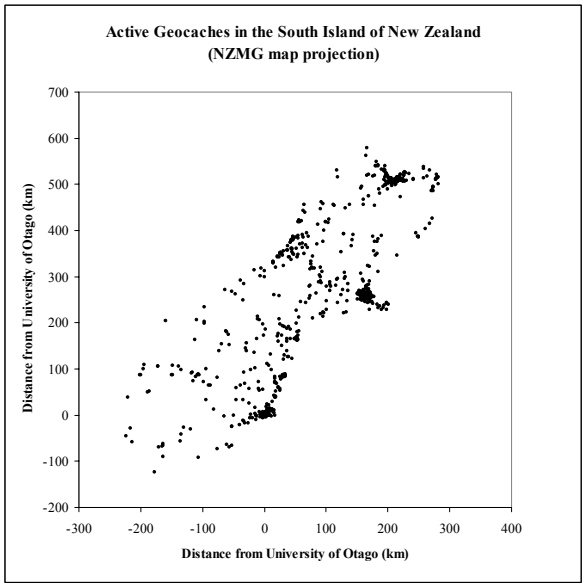



**Figure 1: Geocaches located in the South Island of New Zealand (151,215 km$^2$)**

Experiments were conducted to determine if it is possible to calculate a home-coordinate from the geocaches logged as found by a player.

Actual home-coordinates were solicited from players by posting a message on the geocaching discussion board gps.org.nz. This resulted in 12 replies from South Island players (7.5% of the active players). This set, although small, was used to compute the error in home-coordinate estimate.

### 3.1. Methods

The geocaches in the South Island of New Zealand were trawled from geocaching.com.au (on the 20$^{th}$ of May 2005). This produced a set of 806 geocaches (Figure 1) with 9,271 logs from 390 players who had found those geocaches.

This dataset contained geocaches that, although found by some players, have subsequently been removed from the sport (one of those placed by an author was washed away during a flood). Geocaches marked as unavailable at the time of the trawl were never recommended, but were used to compute user / user similarities and player habits.

Players were divided into two groups, inactive and active.

An inactive player was any player who had not found a single geocache during 2005 (in nearly 5 months), or who had found fewer than 5 in total. The first category includes those who have stopped playing, the second those who have not embraced the sport. Both do not enjoy the sport so building a recommender for them is futile.

After unavailable geocaches and inactive players were removed, there remained 741 geocaches, 160 players, and 8,299 logged finds.

The home-coordinate of each of the 12 players was estimated using five methods.

The geographic mean of all players' known home coordinates is used as a baseline in method *naïve*.

In method *geomean*, the home-coordinate was taken as the geographic mean of the geocaches a player had found.

In method *geomean2sd*, the home-coordinate was computed as the geographic mean of the finds, then those finds outside 2 standard deviations of the mean are removed and the geographic mean recomputed from the remainder set. This method was expected to outperform *geomean* as many players are known to play when on holiday at locations outside their home territory. As these holiday finds are likely to be only a small subset of the total finds of a player, they are likely to fall outside 2 standard deviations of the mean and will, therefore, be filtered out using this method – *geomean2sd* computes the mean from only the remaining finds.

Method *geomean1sd* was computed in the same manner as *geomean2sd*, except those finds outside one standard deviation of the mean were removed before the mean was recomputed.

In method *gaussian*, the smallest north/south-aligned bounding box containing a player's found geocaches was constructed. The bounding box was divided into axis-aligned 1km by 1km squares and at each vertex a Gaussian filter was applied according to equation (1)

$$g(d) = \sum_{c \in C} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{d_c^2}{2\sigma^2}\right)} \qquad (1)$$

where $d_c$ is the Euclidean distance (computed using the NZMG map datum [15]) between the vertex and the geocache, $c$, (from the set of player found geocaches, $C$), and $\sigma$ was set to 50km (a "reasonable" player roaming radius). The vertex with the highest score was considered to be the home-coordinate. This method finds an approximation of the centre of the largest cluster in which the player has found geocaches – it is reasonable to believe this is their home-coordinate.

Each of the four methods was tested for the 12 players for which the home-coordinate was known. The error was computed as the mean Euclidian distance between the predicted coordinate and the player's supplied coordinate. Although this may be subject to over-fitting, the sample is too small to divide into training and evaluation sets.

## 3.2. Results

Figure 2 shows the error for each of the 12 players, and is summarised in Table 1. The two best predictors were *geomean1sd* and *gaussian*.

*Geomean* estimates using those geocaches the player has found, but which are outside their usual playing area. The number of these finds is reduced when those outside 2 standard deviations are discarded. More are discarded using *geomean1sd*. The exception to this is player 2 who is known to have a very large daily roaming area (due to work commitments) and for whom an estimate of a home-coordinate is therefore difficult. Player 8 is known to live in a low population area and to regularly travel to high cache-density areas to play.

| Method | Mean |
|---|---|
| *naïve* | 152.29 |
| *geomean* | 58.27 |
| *geomean2sd* | 51.70 |
| *geomean1sd* | 39.65 |
| *gaussian* | 35.83 |

**Table 1: Mean error in the home-coordinate using each of the estimators**

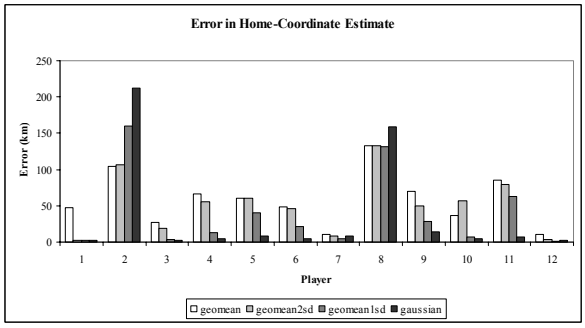



**Figure 2: Error in home-coordinate estimates using each of the given estimators. The two best methods are *geomean1sd* and *gaussian***

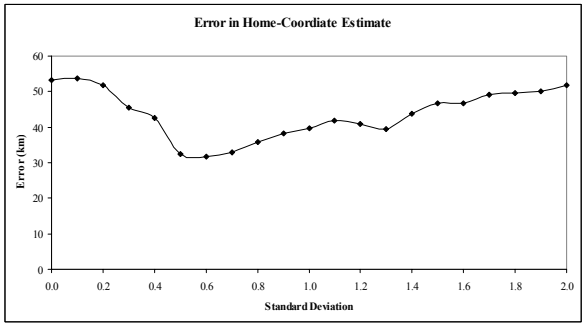



**Figure 3: Error in home-coordinate estimate as a function of geocaches disregarded**

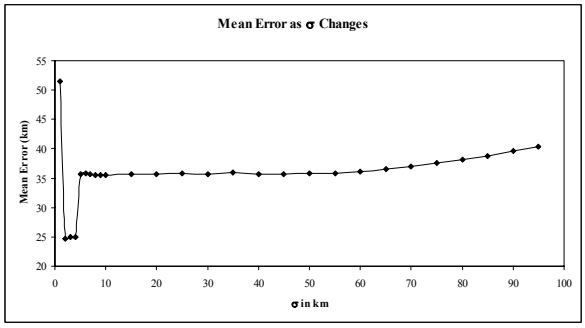



**Figure 4: Error in home-coordinate estimate as a function of the sharpness of the Gaussian filter**

As the error reduces as more finds are discarded, it is pertinent to ask exactly how many standard deviations from the mean should be used as the cut-off. Figure 3 shows the effect of varying this from 0 (using only the closest point to the mean) to 2. The figure shows that as more outliers are discarded the error decreases. Eventually those points on the edge of the player's true centre are discarded and the error increases. The least error occurred when finds outside 0.6 standard deviations were discarded (giving an error of 31.60km).

It is also pertinent to ask how error is affected by a changing standard deviation ($\sigma$) in method *gaussian*. This is plotted in Figure 4 for values in the range 1km to 100km. The trend shows a general decrease in error as the filter increases in sharpness. As the filter decreases in breadth, the score tends to a measure of the density in an ever decreasing sized area. Eventually, it will choose between dense areas, identifying the densest as the player's home – this is evident by the sudden drop at 4km.

As $\sigma$ tends to zero, the score no longer represents a player's ordinary roaming radius, but will find subclusters within that radius, eventually identifying the two closest geocaches the player has found – this is evident in the sudden rise at $\sigma$=1km.

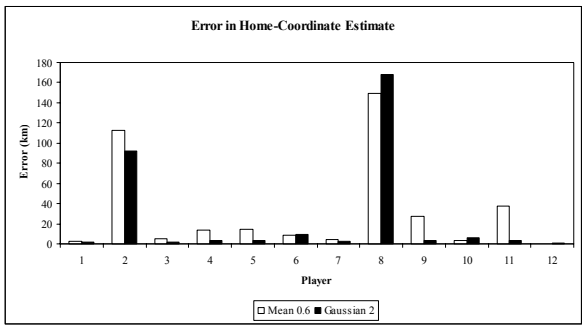



**Figure 5: Comparison of best Gaussian and mean estimators. The Gaussian method has a smaller overall error and is better for most players**

The best score computed using the Gaussian method occurred at $\sigma$=2km (mean error of 24.75km, median of 3.48km) while the best for the mean method occurred at cut-off of 0.6 standard deviations (error of 31.60km, median 11.15km). Figure 5 compares the error using each method for each player. The Gaussian method is better for 8 of the 12 players.

*Gaussian* with $\sigma$=2km is the method we use here on in to estimate the home-coordinate of all players.

# 4. Recommending Good Geocaches

Knowing a player's home-coordinates makes it possible to compute a list of nearby geocaches, however there remains the problem of recommending only those the player will enjoy.

## 4.1. Measuring Performance

Before algorithms can be satisfactorily compared a quantitative comparison method is needed. Upon initial inspection geocaching appears to lend itself to several such methods, but in fact does not.

The player's found list is comparable to a set of relevance judgments. For a given player, those geocaches that have been found are considered relevant, all others non-relevant. The task of the recommender is to recommend the geocaches a player has found based on some analysis. In this way each player is comparable to a query in a traditional information retrieval test collection, consequently mean uninterpolated average precision (MAP) [1] might be used to compute performance.

Using MAP does not take into account the temporal nature of the sport. Just because a player has not found a geocache does not mean they will not find it. For example, all those geocaches placed within a day of the trawl will have been found by very few players, however just a few days later they may have been found by many more.

The recommender could try and match the order the player found the geocaches. A metric such as the normalized distance-based performance measure (NDPM), or the half-life utility metric (see Herlocker *et al*. [6] for details of both) would be used to determine how well the recommended order matched the player's chosen order. However, geocaching is a temporal sport – new geocaches are being added, and old ones decommissioned. A metric trying to match a find order would also have to take into account the life cycle of a geocache.

Metrics that predict user ratings (mean error based metrics [6]) are inappropriate because players cannot rate geocaches.

Each metric measures the performance of a system relative to certain assumed user behaviour (a user model). These assumptions should be stated up-front so it is possible to verify the model – and correct it if erroneous.

For the purpose of this investigation it is assumed that a player finding a geocache is a positive vote for it. The converse, however, is not true (there are no irrelevant items in the collection). This assumption is necessary if the recommender is to be effective when a player moves home-coordinate (just because a player has not found a geocache in Sydney, it does not mean they do not want to find them there if visiting).

It is assumed that at any one moment in time the player chooses what they consider the "best" geocache to find next, and do find that geocache next. This assumption is necessary for two reasons. It makes the player choices discrete and deterministic, and it makes it possible to ignore log entries that log events other than finds (such as a did-not-find in the case of a geocache that has been pilfered).

Most specifically we assume that should one single find be removed from the player's found list then the

very next geocache they choose to find is that very same geocache.

These assumptions turn the spatiotemporal aspects of geocache recommending into a named entity finding problem. As such, the metric of mean reciprocal rank (MRR) is appropriate. Averaging this over each player (the mean of mean reciprocal ranks, (MMRR)) gives a metric that favours each player equally.

We note that McLaughlin and Herlocker [11] recommend precision-based metrics for measuring the performance of collaborative filtering algorithms. We find our problem naturally lends itself to doing so.

For the experiments, the performance of the recommenders is computed by iterating over the list of all players and computing the mean of MRR for each player, according to equation (2)

$$MMRR = \frac{\sum_{p \in P} MRR_p}{|P|} \quad (2)$$

where $P$ is the list of players, $|P|$ is the number of players, and $MRR_p$ is the mean reciprocal rank for player $p$ computed according to equation (3)

$$MRR_p = \frac{\sum_{f \in F_p} RR_{fp}}{|F_p|} \quad (3)$$

where $F_p$ is the list of found geocaches, $|F_p|$ is the number of found geocaches, and $RR_{fp}$ is the reciprocal rank of the geocache in the recommended list, computed according to equation (4)

$$RR_{fp} = \frac{1}{r_{fp}} \quad (4)$$

where $r_{fp}$ is the rank of the given geocache in the list recommended by the system.

By the stated assumptions, two recommender systems can be compared quantitatively using MMRR; but this is not quite enough. A very large positive shift in performance with respect to a single player could have a marked effect on the metric. Exactly this problem is seen in information retrieval experiments where it is now common-place to present the significance of a change using the *t*-test or the Wilcoxon test. Sanderson and Zobel [14] compare the reliability of the two tests on TREC [4] data and suggest the *t*-test is more reliable. MMRR along with significance computed using a one-tailed *t*-test is reported here.

## 4.2. Data Analysis

It seems intuitively obvious that older geocaches have been found more times than newer ones. One would expect a geocache placed in the year 2000 to have been found many more times than one placed last week. To demonstrate this, a plot of age (in 30 day months) against mean number of finds for geocaches of that age is given in Figure 6. There are two points of interest: first, the number of finds is not, in general, a function of age; second, the number of finds is a function of age for some "short time".

The mean and standard deviation of the monthly find rate are 12.88 and 3.75 respectively. Assuming, with reasonable confidence, that any data points above the mean minus one standard deviation are representative of the mean, the intersection of this and the frequency curve will represent the point at which the "short time" ends. This is shown in Figure 7, where the intersection point is between two and three months. For the first three months of the life of a geocache, the number of finds is a function of age, after that, it is not.

As geocaches are geographically dispersed, the chaos around the mean shown in Figure 6 could be caused by geographic isolation of geocaching communities. If this were the case then all geocaches older than three months, ordered by distance from a given point, would show clear peaks in mean find numbers at community centres. In Figure 8 and Figure 9, all geocaches three months and older were ordered by distance from Wellington (North Island) in 10km buckets. Figure 8 shows the number of geocaches in the buckets, while Figure 9 shows the mean number of finds for that bucket. Vertical lines represent (from left to right) Nelson, Christchurch, Timaru, Oamaru, and Dunedin. From visual inspection, there are geocaching centres at large towns, however this has no effect on the mean number of finds of geocaches in the area.

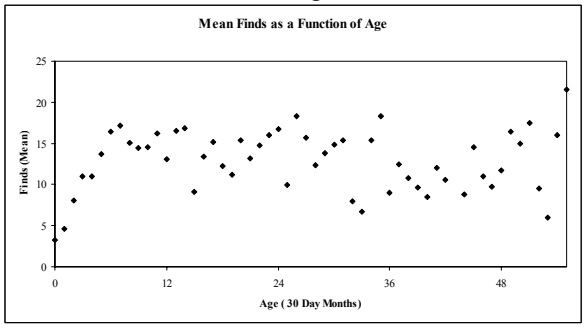



**Figure 6: Mean number of finds for geocaches of the given age (in 30 day months)**

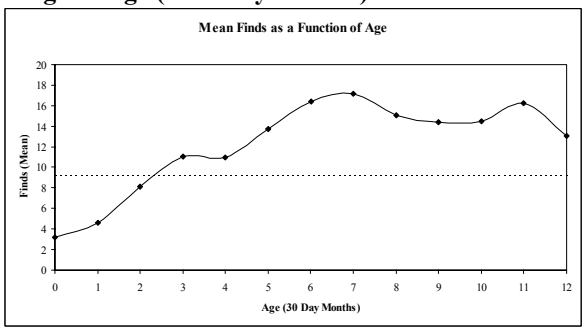



**Figure 7: Mean number of finds for geocaches of the up-to one year. The horizontal line is the mean minus one standard deviation**

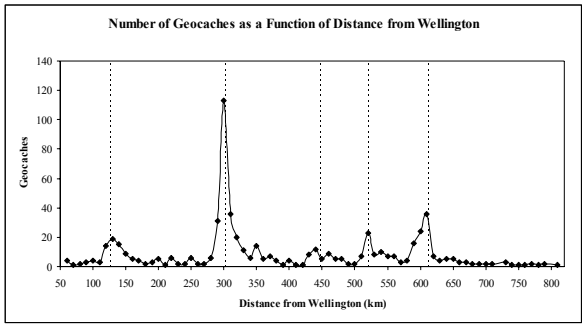



**Figure 8: Number of geocaches as a function of distance from Wellington (10km buckets). Dotted lines are (from left to right) Nelson, Christchurch, Timaru, Oamaru, and Dunedin**

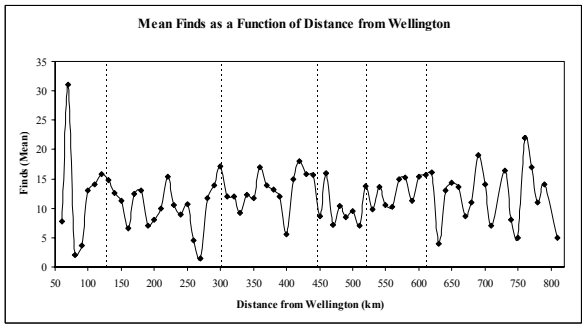



**Figure 9: Mean geocache finds as a function of distance from Wellington (10km buckets)**

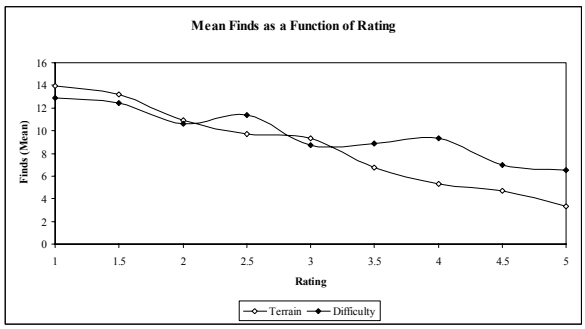



**Figure 10: Mean number of finds for geocaches of the given terrain and difficulty ratings**

If age cannot be used as a predictor of popularity, then what can?

Hiders rate their geocaches on a 5 point scale (including half points) for each of difficulty and terrain (guidelines exist). The easiest receive a score of 1 whereas the most difficult receive a score of 5. The mean number of finds for caches of the given rating is shown in Figure 10. Both show a near linear correlation, as the rating increases the mean number of finds decreases. Other attributes available for analysis include the type of geocache as well as the physical size – these are presented in respectively Figure 11 and Figure 12.

Although geocaching.com has several additional binary attributes (for example if or not climbing gear is needed), these attributes are not present in our data as they are not available on the site we trawled.

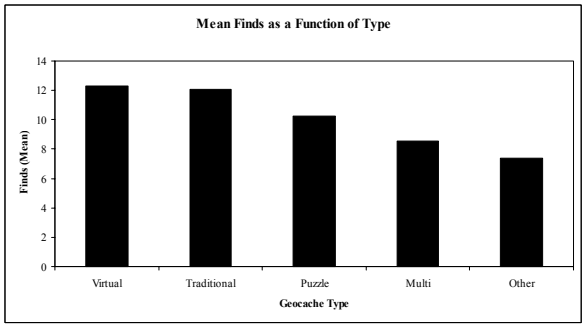



**Figure 11: Mean of finds for geocaches types. The Other category includes Earthcaches (2), Event-caches (5), Letterbox caches (3) and Webcam caches (1)**

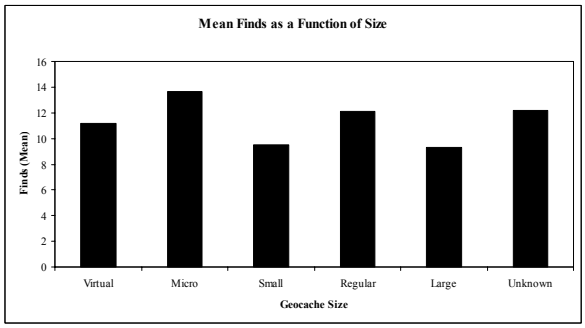



**Figure 12: Mean finds for given sized geocaches**

## 4.3. Possible Recommenders

The data analysis suggests terrain, difficulty, size, and type might be used in recommending geocaches. Additionally the number of finds, and the distance from the home-coordinate might be used. Separately, collaborative filtering techniques might be used. Several techniques were tried.

Geocaching.com orders by *distance* so this is used as a comparative baseline. In this method the next geocache a player chooses to find is the nearest unfound geocache to their home-coordinate.

Assuming the distributions discussed in Section 4.2 are probability distributions (each normalized in to range 0 to 1), the probability that a player will choose to find the given geocache is given by the product of the probabilities for each attribute. All five attributes (terrain difficulty, size, type, and popularity) are used exactly in this way in method *unweighted*.

In method *weighted*, the log of each distribution is weighted by a constant (learned using a genetic algorithm (GA) [7]). This is shown in equation (5)

$$P(R \mid g) = \sum_{k \in \{t,d,s,v,p\}} \left( c_k \times \log P(R \mid k) \right) \qquad (5)$$

where $g$ is the geocache, $k$ predictor ($t$ for the terrain, $d$ for the difficulty, $s$ for the size and $v$ for variety (type), and $p$ for popularity).

A global voting scheme is used as a baseline for collaborative filtering techniques. In method *popularity*, the geocaches are ranked by the number of times found with ties broken by distance.

In method *vote* those *m* players with the closest geocaching behaviour to the given player were found. Similarity was measured using the Tanimoto coefficient (intersection over union) [12]. These players then voted for each geocache they had found, and ties were broken on distance.

Several additional methods were tried (for example, weighted without popularity, sum of probabilities, feature-space similarity, etc.), however none performed as well as the best reported herein.

## 5. Methods

The home-coordinate of each user was computed using the Gaussian method with $\sigma$=2km.

In an iteration of the experiment, a single recommender method is tested. A player is chosen and one geocache find is removed from their found list. The collection statistics (e.g. probability at each difficulty level) are then computed without this find (to remove bias). The *n* closest unfound geocaches to the home-coordinate (excluding those placed by the player) are then ordered according to the recommender method. Finally the MMRR score is computed.

For the collaborative filter the number of similar players, *m*, is varied to achieve the optimal value.

Weights, $c_k$, were learned with a genetic algorithm [7], optimised for $n$=10 nearest geocaches. It was run for 500 generations with a population size of 100, mutation rate of 0.1, single-point crossover rate of 0.6, and reproduction rate of 0.3. Elitism [2] was used with the top 5 individuals carrying over into the next generation (other values were not tried). The experiment was run four times, each had similar results.

The method that works best for the closest few geocaches to the player's home-coordinate might be quite different from the method that works best considering all the geocaches in the whole South Island. To see if such an effect exists, each method was tested on only the closest *n* geocaches to the home-coordinate. Values for *n* varied from 10 to 100 in steps of 10 (representative of pages of results on a web site). Methods that score best at the end of the first page ($n$=10) were considered best as seldom do searchers view past the first page of results [8].

## 6. Results

The results of the experiment are shown in Figure 13. Examining the baseline (*distance*), a clear upwards trend is shown as the number of geocaches included increases. This is because this method is rank-order preserving (with increasing *n*). It is asymptotic because eventually every find is accounted for. The other methods are not rank-order preserving so precision can decrease with increasing *n*.

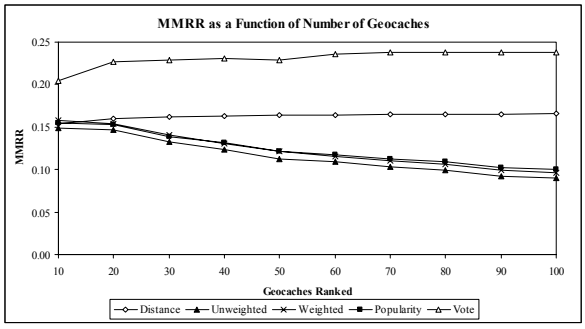



**Figure 13: Performance of each recommender**

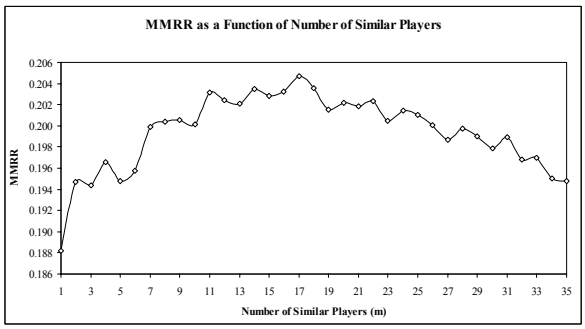



**Figure 14: Effect on MMRR of varying the number of similar players (*m*) used in the *vote* method.**

The best method tried was *vote*, a collaborative filtering scheme in which *m* similar players voted based on their found list. To find the optimal value of *m* the experiment was re-run with varying values for *m*. The result, shown in Figure 14, suggests that for the first page of results ($n$=10) the optimal value is 17 (that used in Figure 13).

In a collaborative filter the performance is known to increase with a decrease in sparsity [5]. In geocaching the sparsity problem does not occur because there are a relatively low number of geocaches in any one geographical area and a relatively large number of players searching for them.

| Popularity | Size | Type | Terrain | Difficulty |
|---|---|---|---|---|
| .424 | .122 | .093 | .351 | .010 |

**Table 2: Weights learned by GA favour mostly popularity and terrain**

The best non-collaborative scheme tried was *weighted* in which the final weights are given in Table 2. Popularity and terrain are favoured most, followed by size, type, and lastly difficulty. Popular geocaches that are easy to get to are, in general, preferred over the others.

Significance computed with a one-tailed *t*-test (at $n$=10 geocaches) show that the improvement of *weighted* over *distance* is not significant ($p$=0.33), but *vote* over *distance* is significant ($p$=0.00).

Increasing the number of geocaches, *n*, used in the ranking does have an effect on the performance of the recommender. Only the collaborative filter and ranking on distance maintained their performance as *n* increased. The *vote* method with $m$=17 scoring the

highest when $n$=100 (MMRR= 0.2380; on average at position 4.2 in the ranked list of results).

The analysis and experiments suggest that if the player has a geocaching history, the best recommender is a collaborative filter using the seventeen most similar players. With no geocaching history it is to use either the weighted method, or distance from home-coordinate

## 7. Conclusions

In the South Island of New Zealand there are 741 active geocaches, mostly located near to high-population centres. A recommender for this sport will help players identify the few geocaches they might enjoy amongst these.

A collection of geocaches, players, and player finds was trawled from the internet. The correct details of player home-coordinate were solicited using an online discussion list.

Several methods of estimating a player's home-coordinate from their logged finds were tested. In the best, computed using a Gaussian filter with $\sigma$=2km, the mean error was 24.75km and the median 3.48km. Although we don't know who the players are, we can identify their home-coordinate. There exists an obvious security issue here (especially should we also be able to identify the players).

Several recommenders are discussed and were tried. Each was tested including varying numbers of geocaches close to the player's home-coordinate. The performance of each was measured using mean of mean reciprocal rank.

The best method tested was a collaborative filter that identified the nearest seventeen players, all of which voted for the geocaches they had found. We recommend using such a method – once the player has a geocaching history. Before then we recommend ordering using a weighted probability method, or by distance.

## References


[1] Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd ACM SIGIR Conference on Information Retrieval*, (pp. 33-40).

[2] De Jong, K. A. (1975). *An analysis of the behavior of a class of genetic adaptive systems.* Unpublished Ph.D., University of Michigan.

[3] Geocaching.com. (2005). Geocaching - the official global gps cache hunt site. Available: http://www.geocaching.com/ [2005, 1 September].

[4] Harman, D. (1993). Overview of the first TREC conference. In *Proceedings of the 16th ACM SIGIR Conference on Information Retrieval*, (pp. 36-47).

[5] Herlocker, J., Konstan, J., & Riedl, J. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information Retrieval,* 5(4), 287-310.

[6] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *Transactions on Information Systems,* 22(1), 5-53.

[7] Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press.

[8] Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th ACM SIGIR Conference on Information Retrieval*, (pp. 154-161).

[9] Kimbo. (2005). A short history of geocaching: May 2000. Available: http://www.guysnamedkim.com/geocache/geocache_history.html [2005, 1 September].

[10] Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M. S., & Duri, S. S. (2001). Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery,* 5(1-2), 11-32.

[11] McLaughlin, M. R., & Herlocker, J. L. (2004). A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *Proceedings of the 27th ACM SIGIR Conference on Information Retrieval*, (pp. 329-336).

[12] Mild, A., & Reutterer, T. (2003). An improved collaborative filtering approach for predicting cross-category purchases based on binary market basket data. *Journal of Retailing and Consumer Services,* 10, 123-133.

[13] Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., & Riedl, J. (2003). Movielens unplugged: Experiences with an occasionally connected recommender system. In *Proceedings of the 8th international conference on intelligent user interfaces*, (pp. 263-266).

[14] Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proceedings of the 28th ACM SIGIR Conference on Information Retrieval*, (pp. 162-169).

[15] Stirling, I. F. (1973). *New zealand map grid* (Technical Circular 1973/32): Department of Lands and Survey, New Zealand.

# Web Searcher Interactions with Multiple Federate Content Collections

*Amanda Spink*
Faculty of IT
Queensland
University of
Technology
QLD 4001 Australia
*ah.spink@qut.edu.au*

*Bernard J. Jansen*
School of IST
The Pennsylvania
State University
PA 16802 USA
*jjansen@ist.psu.edu*

*Chris Blakely*
Infospace, Inc WA
98004 USA
*chris.blakely@Infospace.com*

*Sherry Koshman*
School of Information
Sciences
University of
Pittsburgh
PA 15260 USA
*skoshman@sis.pitt.edu*

**Abstract** Federated content collections are important for providing access to multiple content repositories. Our paper provides preliminary results from a large-scale study of user access to federated content collections via the Dogpile.com Web search engine. We examined differences in searching patterns to various federated content collections by analyzing a subset of queries submitted by searchers on Dogpile.com. Findings include differences in content collection searches. Image and audio searches were longer sessions but shorter queries. Most users view few results pages.

**Keywords** Federated document collection, Web, Dogpile.com

## 1 Introduction

Federated content collection is a content organizing scheme involving multiple repositories of content instead of a central repository. These individual repositories typically have their own store, indexing, and retrieval algorithms. Major Web search engine typically offer tabbed interfaces that permit users to search multiple federated content collections, such as Web documents, images, audio, and video files. Few studies have examined users' access to multiple federated search collections via Web search engines. There are several ongoing projects seeking to build federations of learning content and content repositories [1, 2]. Examination of how people use federated content collections is an important area of document computing research. Our paper provides preliminary results from a large-scale study of user access to federated content collections via the Dogpile.com Web search engine.

The next section of the paper outlines the related studies, followed by the research design and key results from our study.

**Proceedings of the 10th Australasia Document Computing Symposium Sydney, Australia, December 12, 2005.**


## 2 Related studies

Rehak, Dodds and Lannom [3] developed a model and infrastructure for federated learning content repositories. Becarevic and Roantree [4] studied federated multimedia database systems. However, there have been limited studies investigated the effect of federated collections on Web search. Ozmutlu, Spink, Ozmutlu [5] examined the impact of multimedia interface buttons on the Excite search engine, investigating multimedia queries in the general query population prior and after the introduction of radio buttons to search various collections. The researchers reported that the use of radio buttons had decreased the multimedia searches in the general collection. However, the researchers did not examine queries to any of the federated collections.

Jansen, Spink and Pederson [6, 7] compare Web searching characteristics among Web, image, audio, and video content collections on the AltaVista search engine. The researchers report that of the four types of searching, image searching was the most multifaceted task and audio the least complex. The mean terms per query for image searching was notably larger (four terms) than the other categories of searching, which were less than three terms. The session lengths for image searchers were longer than any other type of searching and Boolean usage by image searchers was 28%.

## 3 Research goals

The major research goal of our study was to examine differences in Dogpile.com searching across various federated content collections.

Specific goals were to examine search differences in various federated content collections, including:

1. Session length
2. Query length
3. Number of results pages viewed
4. Use of system assistance

5. Repeat queries

To address these research goals we examined a subset of queries submitted by searchers on Dogpile.com to gain insight into the nature of their search topics.

## 4 Research design

### 4.1 Dogpile.com

Dogpile.com (http://www.Dogpile.com/) is owned by Infospace, a market leader in the meta-search engine business. Dogpile.com is the only meta-search engine during the study period to incorporate the indices of the four leading Web search indices into its search results (i.e., Ask Jeeves, Google, MSN, and Yahoo!). With results from these four Web search engines; Dogpile.com.com leverages one of the most comprehensive content collections on the Web in response to Web searchers' queries. When a searcher submits a query, Dogpile.com simultaneously submits the query to multiple other Web search engines, collecting the results from each Web search engine, removing duplicates results, and aggregating the remaining results into a combined ranked listing using a proprietary algorithm.

Dogpile.com has tabbed indexes for searching the *Web*, *Images*, *Audio*, and *Video*. Dogpile.com also offers query reformulation assistance with query suggestions listed in an "Are You Looking for?" section of the interface. According to Hit Wise[1], Dogpile.com was the 9th most popular Web search engine in 2005 as measured by number of site visits. ComScore Networks[2] reports that in 2005 Dogpile.com had the industry highest visitor-to-searcher conversion rate of 83% (i.e., 83% of the visitors to the Dogpile.com site executed a search).

### 4.2 Data collection

We recorded the records of searcher – system interactions a transaction log that represents a portion of the searches executed on Dogpile.com 6 May 2005. The original general transaction log contained 4,056,374 records. Each record contains three fields:

*User Identification*: an anonymous user code automatically assigned by the Dogpile.com server to identify a particular computer
*Cookie*: anonymous cookie automatically assigned by the Dogpile.com server to identify unique users on a particular computer.
*Time of Day*: measured in hours, minutes, and seconds as recorded by the Dogpile.com.
*Query Terms*: terms exactly as entered by the given user.
*Source*: the content collection that the user selects to search (e.g., Web, Images, Audio, or Video) with Web being the default.
*Feedback*: a binary code denoting whether or not the query was generated by the "Are You Looking for?" query reformulation assistance.

### 4.3 Data analysis

We imported into a relational database the original flat ASCII transaction log file of 4,056,374 records. We generated a unique identifier for each record. We used four fields (*Time of Day*, *User Identification*, *Cookie*, and *Query*) to locate the initial query and then recreate the chronological series of actions in a session.

We define our terminology similar to that used in other Web transaction log studies [8, 9, 10]:
*Term:* series of characters separated by white space or other separator
*Query:* string of terms submitted by a searcher in a given instance
*Repeat query:* query submitted more than once during the data collection period, irrespective of the user.
*Session:* series of queries submitted by a user during one interaction with the Web search engine.
*Session Length*: number of queries submitted by a searcher during a defined period of interaction with the search engine

We also removed all agent and duplicate queries.

## 5 Results

Our paper provides preliminary results from a large-scale study of user access to federated content collections via the Dogpile.com Web search engine.

### 5.1 Content collections

Table 1 shows the usage of each of the five federated Dogpile.com content collections (Web, Images, Audio, Video and News).

[1] Hitwise, 2005. http://www.clickz.com/stats/sectors/search_tools/article.php/3528456.
[2] comScore, 2005 http://www.comscore.com/press/release.asp?press=325.

| Source | Occurrences | Percent |
|---|---|---|
| Web | 1,085,573 | 71.2% |
| Images | 290,571 | 19.07% |
| Audio | 95,118 | 6.2% |
| Video | 48,057 | 3.1% |
| News | 4,474 | 0.29% |
| Total | 1,523,793 | 100% |

Table 1: Use of the Dogpile.com content collections.

Table 1 shows that the Web was the most popular content collection, with more than 71% of all searches being executed again this content collection. Images were the second most popular content collection, followed by the audio, video and news collection.

## 5.2 Session length

Table 2 shows the session length (i.e., number of queries) for queries to the diverse federated content collections.

| Session Length | Web | % | Images | % | Audio | % | Video | % | News | % |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 258204 | 56.843% | 40026 | 51.127% | 10404 | 42.903% | 6741 | 47.462% | 1757 | 69.502% |
| 2 | 77884 | 17.146% | 12420 | 15.865% | 4004 | 16.511% | 2476 | 17.433% | 373 | 14.755% |
| 3 | 40793 | 8.981% | 6328 | 8.083% | 2391 | 9.860% | 1357 | 9.554% | 179 | 7.081% |
| 4 | 24067 | 5.298% | 4087 | 5.220% | 1609 | 6.635% | 881 | 6.203% | 74 | 2.927% |
| 5 | 15341 | 3.377% | 2772 | 3.541% | 1161 | 4.788% | 590 | 4.154% | 47 | 1.859% |
| 6 | 10015 | 2.205% | 2065 | 2.638% | 839 | 3.460% | 412 | 2.901% | 26 | 1.028% |
| 7 | 6839 | 1.506% | 1601 | 2.045% | 667 | 2.751% | 327 | 2.302% | 21 | 0.831% |
| 8 | 4942 | 1.088% | 1202 | 1.535% | 478 | 1.971% | 229 | 1.612% | 13 | 0.514% |
| 9 | 3618 | 0.797% | 1070 | 1.367% | 413 | 1.703% | 205 | 1.443% | 11 | 0.435% |
| 10 | 2581 | 0.568% | 805 | 1.028% | 346 | 1.427% | 158 | 1.112% | 3 | 0.119% |
| 11 | 1873 | 0.412% | 718 | 0.917% | 251 | 1.035% | 115 | 0.810% | 7 | 0.277% |
| 12 | 1506 | 0.332% | 596 | 0.761% | 202 | 0.833% | 105 | 0.739% | 3 | 0.119% |
| 13 | 1130 | 0.249% | 498 | 0.636% | 217 | 0.895% | 73 | 0.514% | 3 | 0.119% |
| 14 | 881 | 0.194% | 438 | 0.559% | 152 | 0.627% | 62 | 0.437% | 3 | 0.119% |
| 15 | 729 | 0.160% | 358 | 0.457% | 118 | 0.487% | 61 | 0.429% | 1 | 0.040% |
| 16 | 609 | 0.134% | 338 | 0.432% | 111 | 0.458% | 50 | 0.352% | 2 | 0.079% |
| 17 | 447 | 0.098% | 282 | 0.360% | 98 | 0.404% | 50 | 0.352% | 1 | 0.040% |
| 18 | 368 | 0.081% | 246 | 0.314% | 80 | 0.330% | 35 | 0.246% | | 0.000% |
| 19 | 326 | 0.072% | 217 | 0.277% | 98 | 0.404% | 27 | 0.190% | | 0.000% |
| 20 | 251 | 0.055% | 203 | 0.259% | 69 | 0.285% | 26 | 0.183% | 1 | 0.040% |

Table 2: Session lengths.

Most users included between one and three queries in their federated content search sessions. Some 50% of users' across the various federated content collections included only one query in their search session. News sessions were shorter and included fewer queries. Audio sessions were longer, but with fewer queries per session.

## 5.3 Query length

Table 3 shows the query length (i.e., number of terms) to the diverse federated content collections.

| Query Length (Terms) | Web | % | Images | % | Audio | % | Video | % | News | % |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 180470 | 16.624% | 74054 | 25.486% | 15470 | 16.264% | 10899 | 22.679% | 744 | 16.629% |
| 2 | 316338 | 29.140% | 122192 | 42.052% | 30008 | 31.548% | 20712 | 43.099% | 1752 | 39.160% |
| 3 | 280473 | 25.836% | 61043 | 21.008% | 20651 | 21.711% | 9930 | 20.663% | 906 | 20.250% |
| 4 | 153570 | 14.146% | 21564 | 7.421% | 13854 | 14.565% | 4116 | 8.565% | 531 | 11.869% |

| Query Length (Terms) | Web | % | Images | % | Audio | % | Video | % | News | % |
|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 77820 | 7.169% | 7643 | 2.630% | 8129 | 8.546% | 1516 | 3.155% | 226 | 5.051% |
| 6 | 38192 | 3.518% | 2577 | 0.887% | 3928 | 4.130% | 533 | 1.109% | 138 | 3.084% |
| 7 | 19192 | 1.768% | 906 | 0.312% | 1749 | 1.839% | 219 | 0.456% | 89 | 1.989% |
| 8 | 10185 | 0.938% | 364 | 0.125% | 829 | 0.872% | 76 | 0.158% | 46 | 1.028% |
| 9 | 5245 | 0.483% | 132 | 0.045% | 312 | 0.328% | 36 | 0.075% | 32 | 0.715% |
| 10 | 2687 | 0.248% | 72 | 0.025% | 112 | 0.118% | 10 | 0.021% | 9 | 0.201% |
| 11 | 1042 | 0.096% | 18 | 0.006% | 53 | 0.056% | 10 | 0.021% | 1 | 0.022% |
| 12 | 290 | 0.027% | 5 | 0.002% | 16 | 0.017% | | 0.000% | | 0.000% |
| 13 | 55 | 0.005% | | 0.000% | 6 | 0.006% | | 0.000% | | 0.000% |
| 14 | 8 | 0.001% | 1 | 0.000% | | 0.000% | | 0.000% | | 0.000% |
| 15 | 2 | 0.000% | | 0.000% | 1 | 0.001% | | 0.000% | | 0.000% |
| 18 | 1 | 0.000% | | 0.000% | | 0.000% | | 0.000% | | 0.000% |
| 24 | 1 | 0.000% | | 0.000% | | 0.000% | | 0.000% | | 0.000% |
| 25 | 2 | 0.000% | | 0.000% | | 0.000% | | 0.000% | | 0.000% |
| | 1085573 | 100.000% | 290571 | 100.000% | 95118 | 100.000% | 48057 | 100.000% | 4474 | 100.000% |

Table 3: Query lengths.

Most queries were between one to three terms per query. Image and audio queries generally included one to two terms. Web, audio and news queries were longer.

## 5.4 Number of results pages viewed

Table 4 shows the number of results pages viewed from the diverse federated content collections.

| Results Pages | Web | % | Images | % | Audio | % | Video | % | News | % |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 781119 | 71.955% | 171869 | 59.149% | 64145 | 67.437% | 32298 | 67.208% | 3123 | 69.788% |
| 2 | 171613 | 15.809% | 53875 | 18.541% | 17853 | 18.769% | 9472 | 19.710% | 905 | 20.223% |
| 3 | 56472 | 5.202% | 37649 | 12.957% | 6730 | 7.075% | 3142 | 6.538% | 240 | 5.363% |
| 4 | 32295 | 2.975% | 12619 | 4.343% | 3097 | 3.256% | 1337 | 2.782% | 110 | 2.458% |
| 5 | 16192 | 1.492% | 5316 | 1.830% | 1274 | 1.339% | 664 | 1.382% | 37 | 0.827% |
| 6 | 9551 | 0.880% | 3741 | 1.287% | 883 | 0.928% | 407 | 0.847% | 27 | 0.603% |
| 7 | 5200 | 0.479% | 1692 | 0.582% | 389 | 0.409% | 230 | 0.479% | 8 | 0.179% |
| 8 | 3621 | 0.334% | 1159 | 0.399% | 270 | 0.284% | 136 | 0.283% | 8 | 0.179% |
| 9 | 2338 | 0.215% | 727 | 0.250% | 138 | 0.145% | 80 | 0.166% | 5 | 0.112% |
| 10 | 1711 | 0.158% | 512 | 0.176% | 105 | 0.110% | 72 | 0.150% | 2 | 0.045% |
| 11 | 1192 | 0.110% | 348 | 0.120% | 66 | 0.069% | 53 | 0.110% | 3 | 0.067% |
| 12 | 854 | 0.079% | 255 | 0.088% | 46 | 0.048% | 39 | 0.081% | 1 | 0.022% |
| 13 | 668 | 0.062% | 172 | 0.059% | 29 | 0.030% | 15 | 0.031% | 1 | 0.022% |
| 14 | 538 | 0.050% | 129 | 0.044% | 27 | 0.028% | 23 | 0.048% | 1 | 0.022% |
| 15 | 397 | 0.037% | 100 | 0.034% | 11 | 0.012% | 19 | 0.040% | 1 | 0.022% |
| 16+ | | | | | | | | | | |
| | 1085568 | 100.0% | 290569 | 100.0% | 95118 | 100.0% | 48057 | 100.0% | 4475 | 100.0% |

Table 4: Viewing of results pages.

Overall, most users' viewed one results page during their search session. More image seeking users also examined second and third page results. Web collection searchers were more likely to view only the first results page.

## 5.5 Use of system assistance

Table 5 shows the use of system assistance when searching the diverse federated content collections. Dogpile.com offers an alternate query re-formulation feature.

| System Assistance | Source | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Web | % | Images | % | Audio | % | Video | % | News | % |
| Yes | 70049 | 6.5% | 44985 | 15.5% | 6236 | 6.6% | 6401 | 13.3% | 455 | 10.2% |
| No | 1015524 | 93.5% | 245586 | 84.5% | 88882 | 93.4% | 41656 | 86.7% | 4019 | 89.8% |
| | 1085573 | 100.0% | 290571 | 100.0% | 95118 | 100.0% | 48057 | 100.0% | 4474 | 100.0% |

Table 5: Use of system assistance.

Across the various content collections, most users did not seek systems' assistance. Interestingly, more users' seeking image and videos sought systems' assistance.

## 5.6 Repeat queries

Table 6 shows the most common repeat queries to the diverse federated content collections.

| | Query | Web | % | Images | % | Audio | % | Video | % | News | % | Total | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | lohan pics | **2586** | **0.238%** | 555 | 0.191% | | | | | | | 3141 | 0.206% |
| 2 | music lyrics | 2436 | 0.224% | | | | | | | | | 2436 | 0.160% |
| 3 | american idol | 1566 | 0.144% | | | | | | | **41** | **0.916%** | 1607 | 0.105% |
| 4 | games | 1240 | 0.114% | | | | | | | | | 1240 | 0.081% |
| 5 | poetry | 1181 | 0.109% | | | | | | | | | 1181 | 0.078% |
| 6 | funny jokes | 1054 | 0.097% | | | | | | | | | 1054 | 0.069% |
| 7 | paris hilton | | | **571** | **0.197%** | | | **203** | **0.422%** | 9 | 0.201% | 783 | 0.051% |
| 8 | google | 694 | 0.064% | | | | | | | 5 | 0.112% | 699 | 0.046% |
| 9 | yahoo | | | | | **676** | **0.711%** | | | | | 676 | 0.044% |
| 10 | ebay | 637 | 0.059% | | | | | | | | | 637 | 0.042% |
| 11 | playstation 2 cheats | 637 | 0.059% | | | | | | | | | 637 | 0.042% |
| 12 | sex | | | 311 | 0.107% | | | 201 | 0.418% | | | 512 | 0.034% |
| 13 | carmen electra | | | 383 | 0.132% | | | 71 | 0.148% | | | 454 | 0.030% |
| 14 | girls | | | 372 | 0.128% | | | 75 | 0.156% | | | 447 | 0.029% |
| 15 | p**** | | | 353 | 0.121% | | | | | | | 353 | 0.023% |
| 16 | britney spears | | | 263 | 0.091% | | | | | | | 263 | 0.017% |
| 17 | eminem | | | | | 243 | 0.255% | | | | | 243 | 0.016% |
| 18 | pamela anderson | | | 214 | 0.074% | | | | | | | 214 | 0.014% |
| 19 | green day | | | | | 209 | 0.220% | | | | | 209 | 0.014% |
| 20 | jennifer lopez | | | 209 | 0.072% | | | | | | | 209 | 0.014% |
| 21 | candy shop | | | | | 177 | 0.186% | | | | | 177 | 0.012% |
| 22 | system of a down | | | | | 174 | 0.183% | | | | | 174 | 0.011% |
| 23 | ludacris | | | | | 163 | 0.171% | | | | | 163 | 0.011% |
| 24 | porn | | | | | | | 135 | 0.281% | | | 135 | 0.009% |
| 25 | hollaback girl | | | | | 133 | 0.140% | | | | | 133 | 0.009% |
| 26 | usher | | | | | 127 | 0.134% | | | | | 127 | 0.008% |
| 27 | lesbians | | | | | | | 86 | 0.179% | | | 86 | 0.006% |
| 28 | funny | | | | | | | 82 | 0.171% | | | 82 | 0.005% |

| | Query | Web | % | Images | % | Audio | % | Video | % | News | % | Total | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | hentai | | | | | | | 78 | 0.162% | | | 78 | 0.005% |
| 30 | jenna jameson | | | | | | | 76 | 0.158% | | | 76 | 0.005% |
| 31 | lesbian | | | | | | | 67 | 0.139% | | | 67 | 0.004% |
| 32 | cdc picaridin | | | | | | | | | 24 | 0.536% | 24 | 0.002% |
| 33 | cdc picaridin sc johnson | | | | | | | | | 24 | 0.536% | 24 | 0.002% |
| 34 | copernic | | | | | | | | | 16 | 0.358% | 16 | 0.001% |
| 35 | kentucky derby | | | | | | | | | 10 | 0.224% | 10 | 0.001% |
| 36 | griswold iowa fire | | | | | | | | | 6 | 0.134% | 6 | 0.000% |
| 37 | "debbie fields" | | | | | | | | | 9 | 0.201% | 9 | 0.001% |
| 38 | 50 cent | | | | | 371 | 0.390% | | | | | 371 | 0.024% |
| 39 | adam long | | | | | | | | | 5 | 0.112% | 5 | 0.000% |
| 40 | akon | | | | | 141 | 0.148% | | | | | 141 | 0.009% |
| 41 | akon lonely | | | | | 119 | 0.125% | | | | | 119 | 0.008% |
| | Total | 12031 | 1.108% | 3231 | 1.112% | 2533 | 2.663% | 1074 | 2.235% | 149 | 3.330% | 19018 | 1.248% |
| | Total (of all queries from this source) | 1,085,573 | 100% | 290,571 | 100% | 95,118 | 100% | 48,057 | 100% | 4,474 | 100% | 1,523,793 | 100.000% |

Table 6. Repeat queries.

Table 6 shows the top ten repeat queries from each content collection. There were nine queries that were the in the top queries from more than one source. Most of these popular people, places, or things.

## 6 Discussion

Our paper provides preliminary results from a large-scale study of user access to federated content collections via the Dogpile.com. Across the federated content collections, there were some differences in users' access. Most searchers accessed the Web collection, followed by the image and audio collections.

Users included between one to three queries in search sessions. Most users' across the various federated content collections entered only one query. News sessions were shorter and included fewer queries. Audio sessions were longer, but with fewer queries per session.

Most users' entered between one and three terms per query. Image and audio queries generally included one to two terms. Web, audio and news queries were longer.

Most searchers examined only the first results page. However, people seeking images examined further results' pages.

Across content collections, most users did not seek systems' assistance. Interestingly, more users' seeking image and videos sought systems' assistance. Image searches were longer and used more terms, Web searches were shorter with fewer queries and viewing fewer results pages. Image and audio searches were longer, including more queries, similar to findings by Jansen, Spink and Pedersen [6], and Spink and Jansen [10].

The nine most frequent queries were for popular people and celebraties, places, or things.

## 7 Conclusion and further research

Our preliminary analysis shows that users' differ in their access to the various content collections. Similar to Web searching overall, most content collection searches are short, and contain few terms and results pages are viewed, except for image searches. We are currently conducting further analysis of the Dogpile.com users and their search processes.

**Acknowledgment** We thank Infospace, Inc for providing the Web search engine data set.

## References


[1] EdNA Online: Education Network Australia http://www.edna.edu.au

[2] Globe (Globe Learning Object Brokered Exchange) http://taste.merlot.org/initiatives/globe.htm

[3] D. R. Rehak, P. Dodds and L. Lannom. A model and infrastructure for federated learning content repositories. In *WWW 2005: International World Wide Web Conference, May 10-14, Chiba, Japan*.

[4] D. Becarevic and M. Roantree. A metadata approach to multimedia database federations. *Information and Software Technology*, Volume 46, Number 3, pages 195-207, 2004.

[5] C. Ozmutlu, Spink and S. Ozmutlu. Multimedia web searching trends: 1997-2001. *Information Processing & Management*, Volume 39, Number 4, pages 611-621, 2003.

[6] B. J. Jansen, A. Spink and J. Pederson. Trend analysis of AltaVista web searching. *Journal of the American Society for Information Science and Technology*, Volume 56, Number 6, pages 559-570, 2005.

[7] B. J. Jansen and A. Spink. An analysis of web searching by European Alltheweb.com users. *Information Processing and Management,* Volume 41, Number 2, pages 361-381, 2005.

[8] B. J. Jansen and U. Pooch. Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*. Volume 52, Number 3, pages 235-246, 2000.

[9] S. Park, H. Bae and J. Lee. End user searching: A web log analysis of NAVER, a Korean Web search engine. *Library & Information Science Research*, Volume 27, Number 2, pages 203-221, 2005.

[10] A. Spink and B. J. Jansen. *Web Search: Public Searching of the Web*. New York: Kluwer, 2004.


.

# Document modelling for customised information delivery

*Shijian LU, Cécile PARIS*

CSIRO ICT Centre
Locked Bag 17,
North Ryde NSW 1670 Australia

*(Shijian.lu, Cecile.paris)@csiro.au*

*Mingfang WU*

CSIRO ICT Centre
Private Bag 33
Clayton South, VIC Australia

*Mingfang.wu@csiro.au*

**Abstract** *As the amount of information available to people multiplies at an increasing speed, it becomes ever more important to deliver information customised to users' specific needs. Natural Language Generation systems coupled with user modeling techniques have been built to address this issue, to produce information that is relevant to the users . A common approach adopted by such systems is an approach based on planning, starting from a discourse (or communicative) goal, and planning the text to be presented to the users. However, these systems are not easy to build and difficult to change by domain experts. One of the problems is that it is hard to specify the plans employed, because they often require knowledge about writing, domain expertise, knowledge of computational linguistics and, finally, knowledge about how to obtain data from the underlying information sources. In this paper, we present our first step to address this problem.*

**Keywords** Document modeling, information retrieval, document generation, personalized documents.

## 1.Introduction

The rapid advancement of information technology has made huge amount information available to more and more people. Increasingly, people depend on the availability of information to achieve their objectives. However, the availability of information does not necessarily translate to productivity gains. In fact, studies have shown that productivity often gets hampered as more and more people are suffering from information overload or fatigue [4]. Indeed, information must be relevant to one's information needs, and it must be easily understandable in order to be useful. That is where contextualised information delivery comes in. Studies have shown that documents tailored to the needs of individual users outperform general purpose documents, e.g., [8, 1, 11, 12].

However, applications delivering customised information are generally expensive to build. Broadly speaking, these applications fall into two categories: template-based and plan-based. There are pros and cons with either approach. Template-based systems are generally easier to construct, but harder to maintain and less flexible, while plan-based systems, with plans of finer level of granularity than typical templates are more flexible and can handle more situations, but require larger overhead to construct [9].

In this paper, we investigate why plan-based tailored document generation systems are difficult to build and report the result of our first step to mitigate the situation. In section 2, we provide some background information and our conception of the problem. Our approach is detailed in Section 3. Before concluding, an example of using this approach is provided in Section 4.

## 2.Background

Despite the fact that customised documents are often more effective than general purpose documents, applications for delivering tailored documents are far less common than they should be. A major reason is that such applications are not easy to develop. Taking plan-based systems as an example, there are at least two reasons why that is the case. First, these systems have typically required extensive semantic knowledge bases [10] which are expensive to craft. Second, the plan operators that underpin the systems' behaviour are difficult to construct.

In CSIRO, we have been developing Myriad [7], a platform for tailored information delivery. The Virtual Document Planner (VDP), its core component, exploits a plan-based approach based on More and Paris [6]. When we developed it, however, we paid particular attention to address the first issue above: we wanted the VDP to produce presentations customised for the user and the situation without the need for a large underlying semantic base. Instead, we wanted to exploit existing technology concerned with retrieving information from existing sources. As a result, the VDP combines discourse planning and document synthesis to gather information through the use of *retrieval services* (in this paper, the notion of *retrieval services* refer to software components which perform information retrieval functions.) that serve as the interface between the two. [2] This alleviates the need for an extensive (usually manually constructed) knowledge base.

The second issue remains: plan operators are typically difficult to write. Before we turn to this problem, we briefly discuss the advantages of using a plan-based approach, as opposed to a template-based one. With a

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.**

```
<operator>
   <id>tellUserAboutStaff<id>
   <effect>(Describe ?staff to ?user)</effect>
   <constraint>(user:isNewStaff ?user)</constraint>
          <nucleus>
                 <value>(inform ?user ?staff homepage)</value>
          </nucleus>
          <satellite>
                 <relation>context</relation>
                 <value>( inform?user ?staff team)</value>
          </satellite>
          <satellite>
                 <relation>context</relation>
                 <value>( inform?user ?staff project)</value>
          </satellite>
          <satellite>
                 <relation>elaboration</relation>
                 <value>( inform?user ?staff informationFromNet)</value>
          </satellite>
   </operator>
```

Figure 1. An example discourse plan operator

plan-based approach, a system starts with a communicative goal, and use discourse plan operators to decompose a high level goal into primitives, see More and Paris [6] for a more detailed description of the process. While doing so, the system builds *a discourse tree*, which is a rich source of information allowing the system to perform a number of reasoning tasks over the generated text.

To illustrate this particular point, let's take the recipe analogy. A recipe typically provides the sequence of steps to be done to produce the dish (i.e., the high-level goal). If something goes wrong (or if the specified ingredient is not available), the person following the recipe cannot reason about what went wrong (or about what other ingredient to use instead of the specified one), as s/he does not know what the purpose of the ingredient is and its role in the overall recipe. The only recourse is to find another recipe for the same dish, hoping this one will succeed (or to find a recipe that does not include that ingredient). Yet if the person understood the role of the ingredients and the steps, s/he may be able to understand what went wrong (or how to substitute another ingredient). Similarly, when producing a multi-sentential document, representing explicitly the intermediary goals and the relationships between the various chunks of information allows the system to understand the role of each element of information and to reason about their role in achieving the main communicative goal. This has been used, for example, to enable a system to participate in a dialogue [6] and to reason about to realise the text on the selected delivery medium [2]. It is because of the resulting discourse tree and the reasoning it enables that we chose the plan-based approach for our platform.

Discourse operators are the plans that tell the VDP how to plan the discourse, and, through their use and expansion, the discourse tree is generated. The plans in the VDP/Myriad include discourse goals and rhetorical relations, the latter based on Rhetorical Structure Theory (RST) [5]. Figure 1 is an example of a discourse operator. It specifies how the discourse goal is to be decomposed into subgoals, thus specifying what content is to be included in the text (at various levels of abstraction – e.g., "describe ?staff to ?user" and inform ?user of specific sub-topics). It also specifies how the text is to be organised (through both the goal decomposition and the use of RST relations, e.g., *context* and *elaboration*, in the figure, which explicitates the relationships between the nucleus and satellites). Operators include constraints which specify the conditions under which the operator is applicable. Finally, operators are of course written in a specific syntax (encoded as XML).

To be competent in writing discourse operators as shown in Figure 1, one needs to possess the following skills.

(1) computational linguistic skills: how to encode a discourse segment in terms of communicative goal and its decomposition (nucleus and satellites) – in particular, understanding of discourse theory, Rhetorical Structure Theory and discourse planning is desirable;
(2) domain knowledge: how to decide under which conditions a plan is applicable and where to get the data;
(3) writing skills: how to write a coherent document appropriate for their audience, and what is the functional role of different parts of the document;
(4) Understand the specific syntax.

Clearly, not many people possess all these skills. As a result, the plans are hard to write for most people, as they

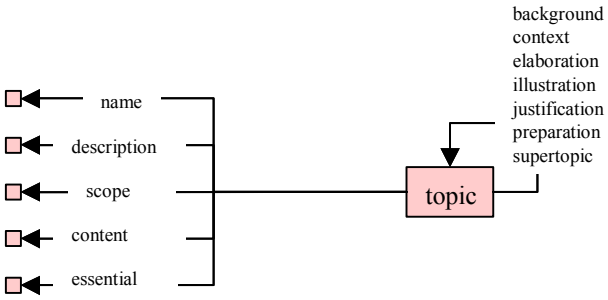



Figure 2. A schematic view of topic

require a lot of expertise, including technical, domain-oriented and writing expertise. Yet it is through plan operators that one specifies the types of text to be generated.

We would like people knowledgeable about the texts required in their domain and with writing skills (so people with writing skills in their domain) to be able to specify these texts, while keeping the advantages of the discourse planning approach, in particular keeping the discourse tree structure that enables further reasoning. To this end, we have started to design a new way to specify the plan operators, to decouple the specification of the structure of the text from the specification of how to retrieve the data, and to provide an abstract way to specify this structure – while still being able to produce the discourse tree.

Our approach is underpinned by three basic constructs: the content structure, the retrieval table and a set of generic operators. The content structure can be seen as a document definition model which is domain dependent. Therefore, it is to be authored by someone who knows how to write texts in their domain. The retrieval table is a registry of retrieval functions available in an application domain. It should be constructed by a software engineer in collaboration with a domain expert, as the data sources themselves are likely to be domain dependent. The set of generic operators is domain independent and have been authored while implementing the approach. They can now be used by different domain applications.

## 3.Modelling document for generation

We have introduced the notion of content structure. In a sense, the content structure is an abstract definition of a dynamic document. A content structure is composed of content nodes and relationships among them. Apart from hierarchical relationship, sibling nodes are related by RST [5]. The content structure can be seen as the blue-print for a tailored document. It (1) defines the rhetorical structure among different chunks of information in a dynamic document; (2) specifies relevant scopes for any particular topic/content node in relation to any contextual models; (3) links retrieval services to the content nodes so that the appropriate data can be retrieved from the underlying data sources.

With the constructs of content structure and retrieval services, we have devised a set of domain independent operators. These domain independent operators, or generic operators, can be used to operate on any content structure for any application domains to generate the desired domain dependent discourse trees. With this approach, there is no need for computational linguistists with domain knowledge to author conventional domain dependent discourse operators for a new application. Instead, only people with domain expertise and writing skills are required to author (domain dependent) content structure. In the following sub-sections, the constructs of content structure, retrieval table and generic operators will be further elaborated.

### 3.1.The Content structure

The content structure, a tree structure with content nodes, is a hierarchical representation of the document to be generated, which could be seen as a hierarchy of topics. At each level of decomposition, there are two types of content nodes (topics): essential and non-essential, essentially mirroring the nucleus/satellite distinction of RST: Essential nodes (nuclei) correspond to primary information that must be included in the text, while information contained in non-essential nodes (satellites) can be secondary or supportive. Again mirroring an RST structure, nodes have to be related with a rhetorical relation. In particular, non-essential nodes have to be related to essential nodes with an RST relationship, e.g., background, context, elaboration, justification, etc. Both types of nodes can be decomposed further.

Figure 2 shows a schematic view of a content node. Each node has a unique name, a textual description, a specification of its scope of applicability, its content proper and an attribute specifying whether it is essential or not.

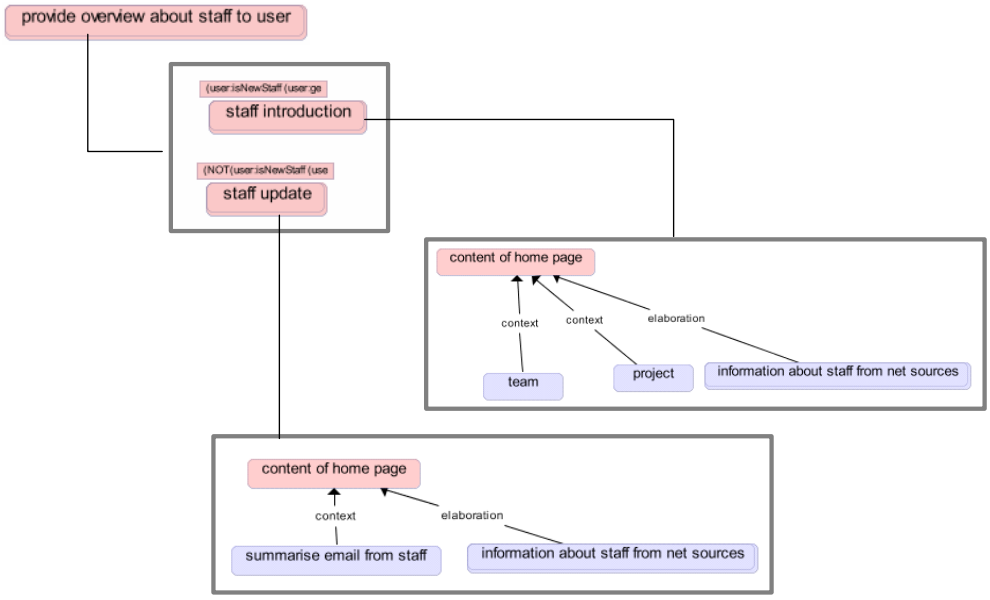



Figure 3. A content structure Fragment

Figure 3 shows a fragment of a content structure. This structure was defined for a new application we are constructing at present: StaffConnector. The system is concerned with providing information about a staff member. This is shown as the top-level node, shaded pink (or darker grey in black&white print), in the top left corner of the picture. The decomposition of the node is shown through the boxes. The author of this content structure decided that the virtual document to be generated in this case (as a web page) consists of either an introduction of the staff member, if the user (reader) is a new staff member him or herself, or an update on the staff member (consisting of a summary of the home page, a summary of email exchanges between this staff member and the user and information from internal sources). Both nodes are then decomposed further.

The scope attribute (shown in the rectangle above the node) defines the applicability or relevance of a topic under certain context. In this example, the "staff introduction" node has a user scope of "user:isNewStaff(user:getCurrentUser))". "user:getCurrentUser" will return the current user from the user model, and "user:isNewStaff(x)" is a Boolean function, which returns TRUE, if x is a new staff, exploiting internal human resource databases. As a result, the node "staff introduction" is only applicable to users who are deemed to be new staff. Scope for a topic is evaluated when the content structure is interpreted by the set of generic plan operators using the constraint mechanism of the plan operators, at runtime.

In essence, the scope provides a simple mechanism to enable "class" level customisation of tailored documents, i.e., inclusion or exclusion of content nodes based on contextual models (which refer to the user, the task, the domain, the discourse history or the environment [7]. (Instance level customisation would refer to different content delivered for different users, but corresponding to the same content node.)

Both nodes are further decomposed: "staff introduction" is further decomposed into the topics: "content of home page", "team" (the staff's position within the organisation hierarchy), "project" (the projects the staff is involved) and "staff on the net" (information related to the staff by searched from the net), where:

- "content of home page" node provides essential information (nucleus, indicated by the pink colour – or darker grey in B&W -- of the node);
- "team" and "project" nodes provide circumstantial information (satellites, shown in blue – or lighter grey in B&W, related to the nucleus by the RST relation called *context*, indicated on the link);
- "information from web sources" provides more information (a satellite, related to its nucleus with the RST relation called *elaboration*).

The content structure thus defines the abstract document structure and how different parts relate to each other. It also defines under what circumstances different parts (nodes) may be applicable (the scope). Finally, it also specifies how to acquire the actual content via *retrieval services* (not shown in the figure). Typically, retrieval services are mapped to the leaf nodes in the content structure. Retrieval services are registered into and managed by the retrieval table, explained in the next section.

As shown in Figure 4, each retrieval service is described by a set of attributes. The <id> field is a unique identifier and, by convention, describes the function of the retrieval service. The <service> field points to the software implementation that will obtain the appropriate data from designated data sources. The <description> field explains what the retrieval service is used for which is what the document designer will see the document design authoring tool. The data sources, which could be files, web sites, databases, and others, are specified in the <access> field. Each service is implemented individually for a specific retrieval purpose, and all retrieval services conform to a common protocol in the Myriad delivery platform.

```
<retrievalTable>
        <retrieval>
                <id>getProject</id>
                <service>GetProject</service>
                <format>xml</format>
                <description> get information about a project </description>
                <access>../../../resources/cmis_org.xml</access>
                <query></query>
                <load>true</load>
                <returnTime></returnTime>
                <returnSize></returnSize>
                <contenttype></contenttype>
                <stylesheet></stylesheet>
        </retrieval>
        <retrieval>
                <id>getHomepage</id>
                <service></service>
                <format>xml</format>
                <description>get staff's home page</description>
                <access></access>
                <query>(retrieval:GetPersonalAttribute (retrieval:getStaff)
                        homepage)</query>
                <load>true</load>
                <returnTime></returnTime>
                <returnSize></returnSize>
                <contenttype>text/html</contenttype>
                <stylesheet></stylesheet>
        </retrieval>
        … …
```

Figure 4. A fragment of retrieval table

## 3.2.Retrieval table

The retrieval table is used to manage retrieval services developed for specific applications. Retrieval services are software components which perform information retrieval functions. There are two types of retrieval services: elementary and composite. Elementary retrieval services directly retrieve needed information, while composite retrieval services are composed of elementary retrieval services or/and other composite retrieval services. Figure 4 shows a fragment of the retrieval table. Here, the first retrieval service, namely, "getProject", is elementary; and the second retrieval service ("getHomepage") is composite.

Composite retrieval services can be defined with the <query> field. In Figure 4, when the "getHomepage" composite retrieval service is called, the retrieval service "getStaff" will be evaluated first, returning the staff id. Then, the retrieval service "GetPersonalAttribute" will be evaluated.

The retrieval table thus contains both elementary and composite services and their definitions. Topics in the content structure refer to retrieval services in the retrieval table. When the content structure gets processed by the generic operators, the generation engine evaluates these retrieval services to acquire information.

```
… …
<operator>
<id>Present0</id>
<description>for composite topics</description>
<effect>(Present ?topic to ?user)</effect>
<constraint>(topic:hasslot ?topic essential )</constraint>
<constraint>(topic:hasslot ?topic normal )</constraint>
<constraint>(not(topic:hasslot ?topic scopeuser))</constraint>
<constraint>(mark name (topic:getslotfiller ?topic name))</constraint>
<nucleus>
  <value>(foreach ?esse (topic:getslotfillers ?topic essential)
        (Present ?esse to ?user))</value>
</nucleus>
<satellite>
  <type>optional</type>
  <relation>background</relation>
  <value>(foreach ?titl (topic:getslotfillers ?topic background)
        (Present ?titl to ?user))</value>
</satellite>
<satellite>
  <type>optional</type>
  <relation>background</relation>
   <value> (foreach ?cont (topic:getslotfillers ?topic context)
         (Present ?cont to ?user))</value>
 </satellite>

… …
<operator>
```

Figure 5. A fragment of a generic plan operator

### 3.3.Generic operators

Having the construct of content structure, we can process it with a set of generic operators and produce a discourse tree akin to the one produced through the conventional discourse operators. The generic operators, starting from the root node of the content structure, perform the following tasks:

(1) evaluate the scope of a node if there is one. The node will not be further processed if result of the evaluation is false;
(2) post appropriate discourse goals by branching out to children nodes;
(3) evaluate any retrieval services that may be attached to a content node, and bind the result in the appropriate structure of the discourse tree.

Figure 5 is an example of generic plan operator. It is worth noting that there is a limited set of generic discourse operators.

## 4. A test application

We tried this new approach on a new application we are developing: StaffConnector. We thus:

(1) developed the retrieval table and associated retrieval services;
(2) authored the domain dependent content structure;

For this application, we developed a simple user model to differentiate new staff from old staff. It is worth noting that the content structure and retrieval table development go hand in hand, even if they can be defined by different people. On the one hand, the content structure sets requirement for what retrieval services are needed. On the other hand, retrieval services determine the content that can be called from the content structure.

To facilitate the authoring of content structure, we have developed a content structure authoring tool, Constructor. This is what was illustrated in Figure 3.

The retrieval service "getHomepage" (Figure 4), which retrieves a specified staff's homepage from ICT Centre's intranet, is attached to the "staff home page" node. Retrieval services are also developed to acquire staff' team information, projects involved in, and information about a specified staff on the net (intranet, extranet, and internet).

Once the retrieval table and content structure are built, they are fed into the VDP together with the set of generic operators. Inside the VDP, the content structure is processed by the generic operators, where applicability scopes get assessed and retrieval services get executed. The outcome is a fully fledged discourse tree, which is then processed with presentation operators. The final outcome is a set of HTML documents which are tailored to the user model. Figure 6 shows the main page of the staff overview application. As it can be seen, there are two panes in the window. The left pane shows the table content view of the document. It provides a global view of the document and can be used to navigate to different sections within the document. By default, the first section is displayed in the right pane. (The layout and presentation are performed by another stage in the planning process.)


Figure 6. The main page for the staff overview application

## 5.Discussion and future work

One of the challenges in developing planning based tailored document generation systems is the issue of authoring discourse operators. The difficulty lies in the requirement of several kinds of expertise simultaneously. To be a competent discourse operator author, one needs to possess knowledge about writing, the application domain and computational linguistics. Consequently, few people are qualified to be able to write discourse operators.

In this paper, we have presented a novel approach to specify the plan operators, decoupling the specification of the structure of the text from the specification of how to retrieve the data, and providing an abstract way to specify this structure – and still being able to produce the discourse tree that is desirable to perform a number of reasoning tasks after the content planning stage. In doing so, we intend to enable people knowledgeable about the texts required in their domain to be able to specify these texts, while keeping the advantages of the discourse planning approach, in particular keeping the discourse tree structure that enables further reasoning. This approach is underpinned by three major constructs, namely, the content structure, the retrieval table and the generic operators.

The content structure is a document definition model which needs to be constructed for every new application. The retrieval table defines retrieval functions for acquiring information from various data sources. Having introduced the content structure, we have developed a finite set of generic operators. With theses generic operators, discourse tree can be generated from any domain dependent content structures. In effect, the issue of discourse operator authoring is transformed into the issue of content structure authoring.

To facilitate the task of content structure authoring, we have built a content structure authoring tool, the Constructor. While we have demonstrated our approach by a simple example, we have also discovered some limitations with our current design. One of the limitations is that the abstract construct of iteration can not be handled. However, we believe that that limitation can be overcome by extending our current modelling constructs. That is what we would like to look into in our next step. Furthermore, we would like to extend our approach to such an extent that it would comfortably handle all possible cases in discourse trees. Another item high on our agenda is to evaluate the usability of our approach.

**Acknowledgements** We wish to thank other members of the group, in particular Andrew Lampert and Akshay Bhurtun.

## References

[1] M.K. Campbell, B.M. DeVellis, V.J. Strecher, A.S. Ammerman, R.F. DeVellis, and R.S. Sandler, (1994). *Improving dietary behavior: The effectiveness of tailored messages in primary care settings*. American Journal of Public Health, 84:783–787.

[2] N. Colineau, C. Paris and M. Wu (2004). *Actionable Information Delivery*. In Revue d'Intelligence Artificielle (RSTI – RIA), Special Issue on Tailored Information Delivery, 18(4), 549-576.

[3] N. Colineau and S. Wan (2001). *Mobile delivery of customised information using Natural Language Generation*. In Monitor (Special Issue on Wireless

Communication Special), 26(3), September-November 2001, 27-31.

[4] A. Edmunds and A. Morris. *The problem of information overload in business organisations: A review of the literature*. International Journal of Information Management, 20(1):17–28, February 2000.

[5] W.C. Mann and S.A. Thompson *"Rhetorical Structure Theory: Toward a functional theory of text organisation"*, In Text 8 (3), 1988, pp. 243-281.

[6] J. Moore and C. Paris (1993) *Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information*. In Journal of Computational Linguistics; 19 (4), December 1993. pp 651 - 694.

[7] C. Paris, M. Wu, K. Vander Linden, M. Post and S. Lu (2004). *Myriad: An Architecture for Contextualized Information Retrieval and Delivery* (2004). In AH2004: International Conference on Adaptive Hypermedia and Adaptive Web-based Systems. August 23-26 2004, The Netherlands. pp.205-214.

[8] C. Paris, M. Wu, A-M Vercoustre, S. Wan, P. Wilkins and R. Wilkinson (2003). *An Empirical Study of the Effect of Coherent and Tailored Document Delivery as an Interface to Organizational Websites*. In The Proceedings of the Adaptive Hypermedia Workshop at the 2003 User Modelling Conference, Pittsburgh, USA, June 22, 2003. pp 133 - 144.

[9] Ehud Reiter. 1995. *NLG vs. templates*. In Proceedings of the 5th European Workshop on Natural Language Generation, Leiden, The Netherlands.

[10] E Reiter and R Dale (2000) *Building Natural Language Generation Systems*. Cambridge University Press.

[11] C.S. Skinner, V.J. Strecher, and H. Hospers, (1994*). Physicians' recommendations for mammography:Do tailored messages make a difference?* American Journal of Public Health, 84:43–49.

[12] V.J. Strecher, M., Kreuter, D.-J. Den Boer, S. Kobrin, H.J. Hospers, and C.S. Skinner. (1994). *The effects of computer-tailored smoking cessation messages in family practice settings.* The Journal of Family Practice, 39:262–270.

# Readability of French as a Foreign Language and its Uses

*Alexandra L. UITDENBOGERD*
School of Computer Science and IT
RMIT
GPO Box 2476V Melbourne Australia
*alu@cs.rmit.edu.au*

**Abstract** *Reading is an important means of foreign language acquisition, particularly for vocabulary. Providing reading material that is of a suitable level of difficulty allows users to acquire vocabulary the most efficiently. Thus an on-line reading material recommender system for language learners requires a readability measure so that the difficulty of texts can be automatically assessed. However, most readability measures were developed for native child speakers of English. In this article I discuss an experiment in readability for learners of French. I conclude that using the average number of words per sentence correlates more closely with human judgements than many commonly available readability measures. I propose a new readability measure for learners of French that have English as their main language, which combines sentence length with the number of words that are similar in both languages (cognates). This measure slightly improves on sentence length for modelling French readability.*

**Keywords** Text readability, Information retrieval

## 1 Introduction

Acquiring sufficient vocabulary to read a foreign language comfortably is an ongoing problem for language learners. Once sufficient grammar is learned the student can make their way through most texts with the aid of a dictionary, but reading more naturally with native-like comprehension remains a dream. Yet, many people need to function at high proficiency in their second, third or even fourth language.

Much research effort has gone into improving language acquisition via reading. Some researchers have concluded that extensive reading at an appropriate level of difficulty is a more efficient method of language acquisition than intensive study of texts [2]. Others have discovered that in order to deduce new words in context requires knowing 95% of the words in the text [4]. This leads to the conclusion that people need to learn a vocabulary of about 5,000 words to be at that level of comfort with normal texts [5].

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 13, 2005.** 

Measures of text difficulty are usually modelled on native children's knowledge of vocabulary and their text comprehension [9]. An example is the Flesch reading ease score available in Microsoft Word. Typical readability measures contain a component representing vocabulary difficulty such as word length and another representing grammatical difficulty such as sentence length. Whilst there are many readability measures developed for English native speakers, there are few for specific foreign languages, and to my knowledge only one that was designed for use across languages. Very few are specifically designed for foreign language students [9].

It is the goal of my research that the acquisition of language through reading can be made more streamlined and efficient through the building of a web text search engine — or really a recommender system — based on readability rather than relevance of topic. In earlier work I examined the issue of readability of French for English speakers [11]. My hypothesis is that current readability measures are not ideal for this purpose as most were developed by using school-age native speakers. Adult or adolescent foreign language learners have different knowledge of a foreign language and different language skills in general. In addition, their previous languages will influence their understanding of the language to be learnt.

In work to be published elsewhere I examine the question of readability of the web, that is, what is the range of readability levels of text on the web. When this is known, it will be clear at what stage the web can be used most efficiently for further language acquisition by reading.

In this paper I once again address the readability of French for English speakers. I scaled up the experiment of my preliminary work [11] by asking people with a range of skill levels in French to rank a set of 10 texts according to difficulty. I also analysed several on-line French books, as well as a corpus of spoken French in terms of vocabulary requirements. It is clear that for some texts a vocabulary of 5,000 would not be quite enough to achieve 95% word knowledge. This gives the learner few stark choices: struggle, read with insufficient understanding, or forget about it. However, there is another option: read selected texts first to build up vocabulary skills before tackling the harder

text. Various software tools can aid in the selecting and sorting of suitable text [4].

## 2 Literature Review

Extensively reading easier texts has been shown to be more effective for language acquisition than intensively reading more difficult texts with the aid of a dictionary [2]. Further corroboration comes from Krantz [10], in whose experiments the students with the strongest language skills gained the most vocabulary through reading a set text. Krantz found that some words can be learned purely through reading, although these words tend to be those that occur frequently in texts, implying a certain level of repetition required [10]. Pioneer of controlled vocabulary-based language teaching Michael West believed that words needed to be encountered initially at least three times before they were absorbed [12]. A later study showed that words need to occur at least five times to be retained (discussed by Ghadirian [4]). Less frequently occurring words are learnt better by looking up in a dictionary when they occur in the text, than just by reading them [10].

Given that learning a language via reading is best achieved with text that is of a suitable reading level, methods of measuring reading level are useful for selecting texts. However, as noted earlier, most of these were developed for native English-speaking schoolchildren. There have been some studies of readability for other languages. Klare mentions that some formulae were tested for English materials to be read by those of a non-English-speaking background [9]. He mentions that Tharp was the first to work on readability for languages other than English. Considerable work on French readability was completed by De Landsheere and his student Henry. More recently Cornaire tested Henry's readability formula for French as a foreign language [3]. However, I'm unaware of any formulae that consider cognates – the words that are recognisably similar to words with the same meaning in the person's native language. For example the word "methode" in French would be a cognate for an English speaker.

Much recent study has been on the use of text corpora to support language learning [13, 6, 7]. Approaches include the study of parallel texts, using concordancers to understand word usage, and the development of targeted vocabularies for learning. A tool that finds web texts based on readability has also been developed [8]. Initially written to find materials of a suitable difficulty level for school-children, the concept can be applied to language learning as well.

## 3 Experiments

In this section I discuss two experiments. The first compares user readability assessments of 10 texts to standard readability measures and factors, as well as across different levels of language skill. The second looks at vocabulary in several on-line French texts.

### 3.1 Relative Readability of Different Texts

The aim of this experiment was to determine how those learning French as a foreign language perceive difficulty of texts. The research questions I raise are:

- What makes a French text easy or difficult for students of French?
- How do current readability measures compare for measuring French readability for students of the language?
- How does French readability for students of French compare to that for native speakers?

#### 3.1.1 Method

In this experiment I wanted to ensure representative samples of various types of French text: native children's books, native adult books, reduced vocabulary books, books designed to have simple grammar, books that intentionally make use of cognates, and books that try to keep both grammar and vocabulary simple. The procedure of selection involved finding the subset from a collection of French books that met the criteria and randomly choosing one book from that subset. In addition I included the draft of a comic book that I have written in which I intentionally restricted the vocabulary to cognates and twelve of the twenty most frequently occurring words found in French newspapers.

A total of fifteen people assessed the selected texts. Table 1 shows the French language skills of the participants. Two participants were native French-speaking adults, and one had spoken French from the age of six. The remainder were students of the Alliance Française in Melbourne. The French skill-level shown is the class that the students were taking at the time of the experiment. Two of the Beginner 2 participants were of Asian descent and may have had English as a second language. The remainder of the participants seemed to have English as their main language. Due to a procedural error two of the participants in the Intermediate 6 class only assessed eight of the ten books.

Each participant was asked to rank the books from easiest to hardest using approximately the first 100 words of the text. Participants varied in the care taken over the task. Some made repeated comparisons. Some flicked through books and made judgements based on this. These rankings were compared with each other as well as with readability measurements.

Approximately the first 100 words (up to the end of the sentence after word 100) were used from each text for readability measurement. The largest number of words used (as counted by the unix utility `wc`), was

| Participant Number | Skill Level | Skill Class |
|---|---|---|
| 1–3 | Beginner 2 | b2/3 |
| 4 | Beginner 3 | b2/3 |
| 5–7 | Beginner 6 | b6/i1 |
| 8 | Intermediate 1 | b6/i1 |
| 9–12 | Intermediate 6 | i6 |
| 13 | native | native |
| 14 | native | native |
| 15 | near-native | native |

Table 1: Language skills of participants in the experiment. Beginner and Intermediate levels refer to those used at the Alliance Française. Beginner 6 is the highest beginner level.

134. The unix `style` utility was applied to each text to gather readability statistics. The statistics included: average words per sentence (WpS), average word length (W len), average number of syllables (Syll), the Kincaid formula (Kinc), the automated readability index (ARI), the Coleman-Liau formula (C.-L.), Flesch reading ease (Fles), the Fog index (Rog), Lix and the SMOG grading. The ARI formula as calculated by `style` is:

$$ARI = 4.71 * Wlen + 0.5 * WpS - 21.43 \quad (1)$$

In addition I manually counted cognates for each text (Cog). A cognate was included if it was either an exact spelling (plus or minus a trailing letter "e"), or a polysyllabic word with an obvious common root and very similar meaning to the English equivalent (eg. compliqué). Repeated cognates were counted. This mainly affected the Gnomeville and Temps des Rèves stories which had some cognates occurring at least 5 times.

I also developed a new measure that combines words per sentence with the cognate count, tuning the constant factor based on the results discussed in the next section.

$$FR = 10 * WpS - Cog \quad (2)$$

In general the cognate count for this formula would be an average per 100 words sampled, but for this experiment the samples of 100-134 were used as a basis for the count.

### 3.1.2 Results

Tables 2, 3, 4, 5 and 6 show the rankings of the texts by participants as well as the mean and standard deviations of these ranks. In table 3 we can see that the standard deviation of the rank is quite low for most books. However, there are a few that are greater than 2. The greatest standard deviation is found amongst the b6/i1 group for the text "La Mission de Slim Kerrigan". This text was adapted to mainly use the 1,000 most frequently occurring words, however, the number of cognates, at least in the first 100 or so words is the lowest in the set of 10 texts. Its sentences were the second longest (See the WpS statistics in Table 7). The greatest difference in average rank across the groups was for "La Grimassouille" a young children's story. The b6/i1 group ranked it three places higher than the other groups (except i6). During the experiment, two of the participants in this group apologised to me that they found this supposedly easy book rather difficult to read. The i6 group also had its largest standard deviation for this book (see Table 6).

The correlation between the different groups was quite high (Table 6), with the lowest correlation being 0.846. (Group i6 was not compared with the others due to the missing data).

Table 8 shows the correlation between standard readability measures and the average ranking for the texts given by each group of participants. In all groups except the native group the best correlated measure was a simple words per sentence count (WpS). For the native group, the ARI measure achieved a slightly higher correlation than words per sentence, and it was the best of the standard readability measures studied. The Flesch score shows a negative correlation as it is a "reading ease" score rather than a reading difficulty score, but was quite weakly correlated. Coleman-Liau gives a negative correlation despite being a reading difficulty score. This may be related to the negative correlation between word length and reading difficulty in this experiment, and indeed between word length and sentence length (-0.45).

Cognates tend to be longer words, having a correlation of 0.65 with word length and 0.68 with syllable count respectively for this collection. The negative correlation between reading difficulty and word length, particularly for non-native participants, may be related to this tendency. The word length effect may be unusually strong in this experiment due to half of the texts being written for students of French that have English as their main language, and the consequent increased use of cognates.

The new measure FR, which combines the cognate count with sentence length, achieved a slightly higher correlation than sentence length alone — except with native French speakers.

The results of the readability experiment suggest that there is a measurable difference in perceived readability between native speakers and learners of the language. The assessment of readability by non-native readers seemed to be much more based on surface features of the language and less on other factors, demonstrated by the higher correlation with sentence length and word length. Comments from two of the native speakers indicated that they took account of the conceptual difficulty of the text in their judgements in addition to other factors. This may account for much of the difference.

| Book | b2 | b2 | b2 | b3 | b6 | b6 | b6 | i1 | i6 | i6 | nat. ad. | nat. ad. | nat. ad. | i6 | i6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 12 | 13 | 14 | 15 | 10 | 11 |
| Les Loisirs | 2 | 3 | 3 | 2 | 1 | 2 | 2 | 2 | 4 | 4 | 1 | 2 | 1 | | |
| Les Miserables | 10 | 8 | 7 | 8 | 3 | 7.5 | 9 | 6 | 7 | 7 | 5 | 10 | 8 | 5 | 7 |
| La Grimassouille | 3 | 2 | 2 | 4 | 7 | 6 | 4 | 5 | 2 | 3 | 3 | 1 | 3 | 7 | 1 |
| Cendrillon | 9 | 7 | 6 | 7 | 10 | 7.5 | 10 | 9 | 9 | 8 | 9 | 7 | 6 | 6 | 5 |
| Les Tours Eiffel | 4 | 6 | 4 | 3 | 4 | 4 | 5 | 3 | 3 | 2 | 4 | 4 | 4 | 1 | 6 |
| Gnomeville: Dragon | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 5 | 2 | | |
| Enfants de Paris | 6 | 4 | 8 | 6 | 6 | 4 | 7 | 7 | 5 | 5 | 6 | 6 | 10 | 2 | 3 |
| Terre des hommes | 7 | 9 | 9 | 9 | 9 | 9 | 8 | 10 | 10 | 10 | 10 | 9 | 9 | 8 | 8 |
| La Mission de Slim | 8 | 10 | 10 | 10 | 8 | 10 | 3 | 8 | 8 | 9 | 8 | 8 | 7 | 3 | 2 |
| Les Temps des Reves | 5 | 5 | 5 | 5 | 5 | 4 | 6 | 4 | 6 | 6 | 7 | 3 | 5 | 4 | 4 |

Table 2: Ranks given to each text by each participant. Where the same rank was given for two or more items the mean rank is allocated to both. For example, the rank 7.5 is given to two items that received equal rank 7 from participant 6, and their is no rank 8.

| Book | all except 10 & 11 | | b2/3 | | b6/i1 | | native | | non-native | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ave | std dev | ave | stdev | ave | stdev | av | std | ave | std |
| Les Loisirs | 2.2 | 1.01 | 2.5 | 0.58 | 1.75 | 0.50 | 1.33 | 0.58 | 2.50 | 0.97 |
| Les Miserables | 7.3 | 1.93 | 8.25 | 1.26 | 6.38 | 2.56 | 7.67 | 2.52 | 7.25 | 1.87 |
| La Grimassouille | 3.5 | 1.71 | 2.75 | 0.96 | 5.50 | 1.29 | 2.33 | 1.15 | 3.80 | 1.75 |
| Cendrillon | 8.0 | 1.39 | 7.25 | 1.26 | 9.13 | 1.18 | 7.33 | 1.53 | 8.25 | 1.36 |
| Les Tours Eiffel | 3.8 | 0.99 | 4.25 | 1.26 | 4.00 | 0.82 | 4.00 | 0.00 | 3.80 | 1.14 |
| Gnomeville: Dragon | 1.5 | 1.13 | 1 | 0.00 | 1.25 | 0.50 | 3.00 | 1.73 | 1.10 | 0.32 |
| Enfants de Paris | 6.2 | 1.63 | 6 | 1.63 | 6.00 | 1.41 | 7.33 | 2.31 | 5.80 | 1.32 |
| Terre des hommes | 9.1 | 0.86 | 8.5 | 1.00 | 9.00 | 0.82 | 9.33 | 0.58 | 9.00 | 0.94 |
| La Mission de Slim | 8.2 | 1.88 | 9.5 | 1.00 | 7.25 | 2.99 | 7.67 | 0.58 | 8.40 | 2.12 |
| Le Temps des Reves | 5.1 | 1.04 | 5 | 0.00 | 4.75 | 0.96 | 5.00 | 2.00 | 5.10 | 0.74 |

Table 3: Average and standard deviation of ranks across all participants (except 10 and 11), and across each group.

| All except 10 & 11 | | Native | | Non-native | |
|---|---|---|---|---|---|
| | mean | | mean | | mean |
| Gnomeville: Dragon | 1.5 | Les Loisirs | 1.33 | Gnomeville: Dragon | 1.10 |
| Les Loisirs | 2.2 | La Grimassouille | 2.33 | Les Loisirs | 2.50 |
| La Grimassouille | 3.5 | Gnomeville: Dragon | 3.00 | La Grimassouille | 3.80 |
| Les Tours Eiffel | 3.8 | Les Tours Eiffel | 4.00 | Les Tours Eiffel | 3.80 |
| Le Temps des Reves | 5.1 | Le Temps des Reves | 5.00 | Le Temps des Reves | 5.10 |
| Enfants de Paris | 6.2 | Cendrillon | 7.33 | Enfants de Paris | 5.80 |
| Les Miserables | 7.3 | Enfants de Paris | 7.33 | Les Miserables | 7.25 |
| Cendrillon | 8.0 | Les Miserables | 7.67 | Cendrillon | 8.25 |
| La Mission de Slim | 8.2 | La Mission de Slim | 7.67 | La Mission de Slim | 8.40 |
| Terre des hommes | 9.1 | Terre des hommes | 9.33 | Terre des hommes | 9.00 |
| | std dev | | std dev | | std dev |
| Terre des hommes | 0.86 | Les Tours Eiffel | 0.00 | Gnomeville: Dragon | 0.32 |
| Les Tours Eiffel | 0.99 | La Mission de Slim | 0.58 | Le Temps des Reves | 0.74 |
| Les Loisirs | 1.01 | Les Loisirs | 0.58 | Terre des hommes | 0.94 |
| Le Temps des Reves | 1.04 | Terre des hommes | 0.58 | Les Loisirs | 0.97 |
| Gnomeville: Dragon | 1.13 | La Grimassouille | 1.15 | Les Tours Eiffel | 1.14 |
| Cendrillon | 1.39 | Cendrillon | 1.53 | Enfants de Paris | 1.32 |
| Enfants de Paris | 1.63 | Gnomeville: Dragon | 1.73 | Cendrillon | 1.36 |
| La Grimassouille | 1.71 | Le Temps des Reves | 2.00 | La Grimassouille | 1.75 |
| La Mission de Slim | 1.88 | Enfants de Paris | 2.31 | Les Miserables | 1.87 |
| Les Miserables | 1.93 | Les Miserables | 2.52 | La Mission de Slim | 2.12 |

Table 4: Sorted mean and standard deviation of ranks for each group

| | b2/3 | b6/i1 | i6 | nat |
|---|---|---|---|---|
| b2/3 | 1 | 0.85 | | 0.92 |
| b6/i1 | | 1 | | 0.85 |
| i6 | | | | |
| nat | | | | 1 |

Table 5: Correlation between different groups

| Book | i6 | | | | average | stddev |
|---|---|---|---|---|---|---|
| | 9 | 12 | 10 | 11 | | |
| Les Tours Eiffel | 2 | 1 | 1 | 6 | 2.5 | 2.38 |
| La Grimassouille | 1 | 2 | 7 | 1 | 2.75 | 2.87 |
| Enfants de Paris | 3 | 3 | 2 | 3 | 2.75 | 0.50 |
| Le Temps des Reves | 4 | 4 | 4 | 4 | 4 | 0.00 |
| La Mission de Slim | 6 | 7 | 3 | 2 | 4.5 | 2.38 |
| Les Miserables | 5 | 5 | 5 | 7 | 5.5 | 1.00 |
| Cendrillon | 7 | 6 | 6 | 5 | 6 | 0.82 |
| Terre des hommes | 8 | 8 | 8 | 8 | 8 | 0.00 |

Table 6: Rankings of the eight texts by the i6 group of participants.

| Book | WpS | W len. | Syll. | Kinc. | ARI | C.-L. | Fles. | Fog | Lix | SMOG | Cog |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cendrillon | 20.0 | 3.74 | 1.29 | 7.4 | 6.2 | 6.2 | 77.6 | 9.5 | 28.8 | 7.7 | 7 |
| Enfants de Paris | 8.8 | 4.28 | 1.46 | 5.1 | 3.2 | 9.4 | 74.2 | 7.7 | 24.9 | 8.2 | 11 |
| Gnomeville: Dragon | 4 | 4.44 | 1.48 | 3.4 | 1.4 | 10.3 | 77.4 | 5.3 | 28.1 | 6.3 | 42 |
| Grimassouile | 9.2 | 4.23 | 1.36 | 4.1 | 3.1 | 9.1 | 82.4 | 5.1 | 25.5 | 6.2 | 11 |
| Les Loisirs | 4.8 | 3.47 | 1.25 | 1.0 | -2.7 | 4.6 | 96.6 | 3.1 | 12.4 | 5.0 | 6 |
| Les Mis. (adapted) | 11.6 | 4.1 | 1.28 | 4.0 | 3.6 | 8.3 | 86.9 | 6.5 | 28.9 | 7.1 | 7 |
| Les Tours Eiffel | 13.0 | 3.92 | 1.31 | 4.9 | 3.5 | 7.3 | 83.0 | 7.9 | 27.5 | 8.2 | 15 |
| Slim Kerrigan | 17.7 | 3.67 | 1.15 | 4.9 | 4.7 | 5.8 | 9.5 | 7.8 | 21.4 | 6.2 | 5 |
| Les Temps des Reves | 13.8 | 3.83 | 1.25 | 4.5 | 3.5 | 6.7 | 87.5 | 8.4 | 24.7 | 8.5 | 12 |
| Terre des Hommes | 15.5 | 3.78 | 1.24 | 5.1 | 4.1 | 6.5 | 86.0 | 6.8 | 25.2 | 5.7 | 12 |

Table 7: Readability statistics for the text samples used in the experiment.

| Group | WpS | W len. | Syll. | Kinc. | ARI | C.-L. | Fles. | Fog | Lix | SMOG | Cog | FR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| all | 0.83 | -0.31 | -0.51 | 0.67 | 0.73 | -0.30 | -0.39 | 0.63 | 0.27 | 0.15 | -0.56 | 0.84 |
| b2/3 | 0.79 | -0.36 | -0.60 | 0.56 | 0.66 | -0.36 | -0.48 | 0.59 | 0.18 | 0.13 | -0.61 | 0.82 |
| b6/i1 | 0.85 | -0.23 | -0.41 | 0.78 | 0.81 | -0.22 | -0.29 | 0.64 | 0.36 | 0.17 | -0.56 | 0.86 |
| native | 0.70 | -0.10 | -0.30 | 0.67 | 0.72 | -0.09 | -0.35 | 0.66 | 0.40 | 0.24 | -0.33 | 0.69 |
| non-native | 0.85 | -0.36 | -0.57 | 0.66 | 0.72 | -0.36 | -0.39 | 0.62 | 0.22 | 0.12 | -0.61 | 0.87 |

Table 8: Correlation between readability measures and mean user rankings.

### 3.2 Vocabulary of French Texts

In this experiment I examined the vocabulary size required to be able to understand 95% of the text of several books and corpora. The texts examined were the French bible, a corpus of spoken French, Consuelo by nineteenth century author George Sand, and Le Petit Prince by Saint-Exupéry. Figure 1 shows the vocabulary (types) required for different portions of the given text sources.

The vocabulary required for the children's book Le Petit Prince is comfortably less than 1500 words, however Consuelo requires somewhat more than 5,000. The spoken French corpus requires a vocabulary that is less than 2,000, and the French bible needs about 4,500 words. This suggests that children's books and conversational vocabulary may be achievable, but that long adult texts will be a challenge. The figure 5,000 for required vocabulary size seems to be supported in these examples, but obviously this experiment is somewhat small in scale for any extrapolation to other texts. It also emphasises that vocabulary requirements grow with the text size, making shorter texts a good choice for learners. This is reflected in current practice, as most reading books for learners of a language are quite short (for example, those published by Hachette and Oxford for French and English learners respectively).

## 4 Conclusions

With the aim of providing an on-line reading recommender for language learners based on readability, I've explored various factors that affect readability. In a study of vocabulary size of French text, my results confirm the previously cited figure of 5,000 as a required vocabulary size for ease of reading. Smaller vocabularies are likely to be required for reading children's books and for general conversation, but even these vocabularies grow with the amount of text.

In my study into French readability for English native speakers, participants ranked a set of texts. These were compared to some readily available readability measures as well as the number of cognates. Amongst the set of measures, the best was a simple average of the number of words per sentence — a similar finding to my earlier work [11]. Combining this with a cognate count gave slightly better results.

There were slight differences between native and non-native speaker assessments of texts but there was stronger correlation between these groups than there was between most group's rankings and the readability measures.

It may be thought that other factors, such as familiarity with the topic discussed, or the story outline would be important for readability. While there is evidence that a rich reading environment that incorporates images aids comprehension [1], storyline familiarity was clearly not a strong factor in this experiment, since both Cinderella (Cendrillon) and Les Miserables were rated as quite difficult by participants in this study. However, two native-speaking participants commented that their readability assessments incorporated the conceptual difficulty of the texts — a factor that would be difficult to measure in text.

The number of cognates in the text was reasonably highly correlated with readability, and when combined with sentence length predicts readability well, as assessed by English-speaking learners of French. This work did not clearly distinguish the relative importance of cognates and sentence length, however, as there was only one text with a markedly different number of cognates and it also had the shortest average number of words per sentence. Future work should probably include the exploration of this aspect of readability.

If cognates are important for measuring French readability for English speakers, then for it to be useful for a text recommender, automated means of identifying cognates in text will need to be developed. This is expected to be the next step in this research project.

## 5 Acknowledgements

I thank the Alliance Française of Melbourne for their assistance with experiments. I also thank Betsy Kerr for providing the French spoken language corpus.

## References

[1] K. Al-Seghayer. The effect of multimedia annotation modes on l2 vocabulary acquisition: a comparative study. *Language Learning and Technology*, Volume 5, Number 1, pages 202–232, January 2001.

[2] T. Bell. Extensive reading: speed and comprehension. *The Reading Matrix*, Volume 1, Number 1, April 2001.

[3] C. M. Cornaire. La lisibilité: Essai d'application de la formule courte d'henry, au français langue étrangère. *Canadian Modern Language Review*, Volume 44, Number 2, pages 261–273, January 1988.

[4] S. Ghadirian. Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning and Technology*, Volume 6, Number 1, pages 147–164, January 2002.

[5] P. J. M. Groot. Computer-assisted second language vocabulary acquisition. *Language Learning and Technology*, Volume 4, Number 1, pages 60–81, May 2000.

[6] S. Hunston. *Corpora in applied linguistics*. The Cambridge Applied Linguistics Series. Cambridge University Press, Cambridge, 2002.

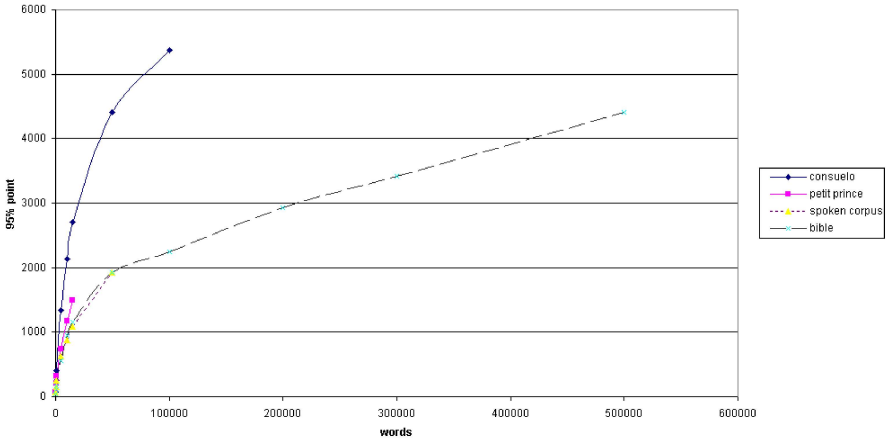



Figure 1: Graph of the 95% vocabulary point for four different French texts. The horizontal axis refers to the number of words from the start of the text that were included in the calculation.

[7] E. St. John. A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning and Technology*, Volume 5, Number 3, pages 185–203, September 2001.

[8] I. R. Katz and M. I. Bauer. Sourcefinder: Course preparation via linguistically targeted web search. *Educational Technology and Society*, Volume 4, Number 3, pages 45–49, 2001.

[9] G. R. Klare. Assessing readability. *Reading Research Quarterly*, Volume X, pages 62–102, 1974.

[10] G. Krantz. *Learning vocabulary in a foreign language: a study of reading strategies*. Ph.D. thesis, University of Göteborg, Sweden, 1991.

[11] A. L. Uitdenbogerd. Using the web as a source of graded reading material for language acquisition. In W. Zhou, P. Nicholson, B. Corbitt and J. Fong (editors), *International Conference on Web-based Learning*, Volume 2783 of *Lecture Notes in Computer Science*, pages 423–432, Melbourne, Australia, August 2003. Springer.

[12] M. West. The construction of reading material for teaching a foreign language. *Dacca University Bulletin*, Number 13, 1927. Republished as part of "Teaching English as a foreign language, 1912 – 1936: Pioneers of ELT, Volume V: Towards Carnegie", R. Smith editor.

[13] D. Wible, C.-H. Kuo, F.-Y. Chien and C.C. Wang. Toward automating a personalized concordancer for datadriven learning: a lexical difficulty filter for language learners. In B. Ketteman and G. Marko (editors), *Conference on Teaching and Language Corpora*, Volume 4, pages 147–154, Graz, July 2000. Rodopi.

# In Search of Reliable Retrieval Experiments

*William Webber and Alistair Moffat*

Department of Computer Science and Software Engineering
The University of Melbourne
Victoria, Australia 3010

{*wew, alistair*}*@cs.mu.oz.au*

**Abstract** *There are several ways in which an "improved" technique for solving some computational problem can be defended: by mathematical argument; by simulation; and by experimental validation. Each of these has risks. In this paper we describe some of the issues that arose during an experimental validation of architectures for distributed text query evaluation, and the approaches that were taken to resolve them. In particular, collections and clusters must be scaled in a way that maximizes comparability between different data sizes; query sets must be appropriate to the target collection; and hardware issues such as file placement on disk must also be considered. Our intention is to report on our experience in a practical sense, and thereby assist others to avoid the same problems.*

## 1 Introduction

There are several ways in which an "improved" technique for solving some computational problem can be defended: by mathematical argument; by simulation; and by experimental validation. Each of these has risks. For example, a mathematical analysis might be erroneous, or might apply only for impossibly large problem domains, or might be predicated on a model of computation that does not reflect actual computer hardware. Similarly, a simulation might be flawed because it fails to account for some aspect of the real-world behavior that is being modeled.

In this paper we describe some of the issues that arose during an experimental validation of architectures for distributed text query evaluation, and the approaches that were taken to resolve them. The issues discussed are "real", in the sense that each of them turned out to be a major impediment to accurate measurement in a set of experiments that we were running, but had not been anticipated at the time the experiments were initially planned. In particular, we found that collections and clusters must be scaled in a way that maximizes comparability between different data sizes; query sets must be appropriate to the target collection; and hardware issues such as file placement on disk must also be considered.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

Our intention is to report on our experience in a practical "warts and all" sense. In our paper describing the new query distribution technique (now being reviewed [Moffat et al., 2005]), we simply stated how the experiments had been run, as if that *modus operandi* had always been the intention. The reality is somewhat different, and our experimental validation took more than a year longer than originally planned, and required a complete rethink of both the software we were testing and also what it was we were planning to measure. We hope that in admitting to our experiences we can inform others planning experimental validations, and thereby assist them to avoid the same problems. Readers who benefit from this commentary might also be interested in the work of Zobel et al. [1996].

## 2 The challenge

Two standard architectures for distributed text query evaluation are described in the literature. In *document partitioning*, each node in the cluster indexes a different subset of the collection's documents. The central receptionist distributes each query to all of the nodes; and each node evaluates the query against its local index, returning the results to the receptionist. The receptionist merges these results and returns them to the user. In *term partitioning*, each node handles a subset of the index's vocabulary. Queries are evaluated centrally by the receptionist, using index information supplied by the relevant nodes. Previous experimental investigations [Tomasic and García-Molina, 1993, Jeong and Omiecinski, 1995, Ribeiro-Neto and Barbosa, 1998, Cahoon et al., 2000, Badue et al., 2001] had been inconclusive, and we had been engaged in debate as to which method provided superior performance. A new evaluation strategy – denoted *pipelining* – emerged out of that debate, and late in 2003 we set in motion plans to test all three strategies. The pipelined mechanism again makes use of a term-partitioned index, but the evaluation state is shipped between cluster nodes, and each node holding information about a query participates in the evaluation of that query.

A key goal of the experiments was that the testing should be under conditions approximating that of a large-scale, real-world search engine, in accordance

with the position we had already argued for previously [Moffat and Zobel, 2004]. The results should not only probe the potential of our new architecture; they should also conclusively demonstrate the relative merits of document and term partitioning, and settle the arguments we had engaged in.

The uni-processing `Zettair` text retrieval engine (available from `http://www.seg.rmit.edu.au`) was used as a basis for the experiments, with code added to implement each of the distributed architectures. We wanted to explore scalability in two directions: as the number of nodes in the cluster increased; and as the size of the collection grew. Cluster scalability would be investigated by performing runs on one, two, four, and eight machines, and comparing query throughput rates. Similarly, collection scalability would be provided through the use of standard TREC collections of different sizes, `GOV`, `wt100g`, and `GOV2`, being respectively $18$ GB, $100$ GB, and $426$ GB in size. Using these three data sets also had the (as it turned out, specious) attraction of testing the systems on different types of collection. To emphasize the practical nature of the experiments, we chose a convenient real-world query log, the public `Excite97` log. Finally, to eliminate startup effects, the first twenty thousand queries were taken from the log, but timings were taken against the second group of ten thousand rather than the whole set.

With this setup in place, we ran our experiments, got interesting results that made the pipelining strategy look good (and simultaneously exposed the term-partitioning approach as being hopeless), wrote everything up, submitted a paper, and got rejected. In retrospect, most of the reasons given by the referees were appropriate, but as always in such a situation, we felt somewhat disheartened.

Other tasks then intervened; when we returned to the investigation a couple of months later, it was with new hardware, a new version of the `Zettair` engine, and with added instrumentation in the software to allow more data to be collected. These changes incorporated, the experiments were re-run, the presentation revised to include the new data, and the paper re-submitted (to a different venue).

Two weeks later a casual corridor conversation led to the quite shocking realization that there was a major problem with the experiments (described in more detail below), and we withdrew the second submission before – we hope – too much editorial and reviewing effort had been invested in it.

Determined to make our third attempt the last, we re-thought and redesigned the experiments, and checked all of the outputs carefully. But there were several more iterations of design needed, of both software and experiment, before we recently submitted (again) the paper describing the pipelined approach to distributed retrieval. In all, this saga took nearly two years from conception to completion, involved a surprisingly steep learning curve, and taught us several hard lessons about designing and running experiments on distributed systems. Our purpose in writing this paper is to describe the path that was followed, the ways in which flaws in the initial experiments were uncovered, and then the ways they were eventually (we believe) rectified.

## 3 Homogeneous data

The first issue for reflection was the decision to use three different collections to explore the manner in which the distribution architectures scaled with collection size. At face value, use of three collections of different sizes allows both exploration of scale effects, and also exploration of different types of data. In particular, `GOV` and `GOV2` are both derived from US government web-sites, albeit a couple of years apart (the former in 2002, the latter in 2004). On the other hand, `wt100g` is quite different. It was crawled from the general web in 1997, and was (at least by intent) restricted to HTML pages, whereas the government crawls include many long PDF files. That is, as well as being different in size, the three collections had different subject matter and document length.

Viewing these additional differences as a chance to kill two investigative birds with one experimental stone was misguided. A key tenet of experimental investigations is to know which attributes are being varied, and which attributes should be fixed, perhaps temporarily, or perhaps permanently. An obvious corollary is to then ensure that in any given experiment only one parameter is being varied, thereby "keeping it simple, stupid". By simultaneously mixing data types and collection sizes in our first set of experiments, we were unable to distinguish between alternative possible effects, and were led to erroneous conclusions.

In the subsequent experiments the two smaller collections were dropped. Instead, fractional collections were created by extracting slices out of `GOV2`. This had the benefit of also allowing for sub-collections to be sized in the exact ratios required, important when investigating the effect of (for instance) doubling both the number of nodes and the size of the collection.

Sub-collection extraction does, however, need to be undertaken with care. It would be wrong, for instance, to create a half collection by simply taking the first half of the documents in the `GOV2` repository. This would ignore the way that web crawls proceed, starting from top-level seed documents, and proceeding to deeper, more obscure ones, and would not have created a sub-collection that was homogeneous with respect to the main one. This is particularly the case with `GOV2`, where all the PDFs are stored at the end. Instead, to make a $1/n$th sub-collection we selected every $n$th file of the $27{,}204$ files making up the `GOV2` collection.

We have chosen to present the main themes of this paper as a sequence of "morals". We begin with this simple one, blindingly obvious, but, nevertheless, one we lost sight of:

*Moral:* When testing a system, only vary the things that need to vary. Fix everything else.

## 4 Appropriate query set

Our initial experiments used the Excite97 query log [Jansen et al., 1998, Spink et al., 2001], which had the benefits of being publicly available and widely known. More importantly, it was attractive because it was "real".

However, we subsequently found that the Excite97 log was a poor fit with the GOV2 collection which we used for the main experiments, for two reasons: first, it was collected in 1997, whereas GOV2 was crawled in 2004; and secondly, it is from a whole-of-web search engine, whereas the GOV2 collection is confined to US government web pages and documents. The mismatch means that many of the queries refer to information and resources that government web sites are unlikely to provide. The five most popular queries in Excite97 are "sex", "yahoo", "chat", "playboy", and "porn"; the most popular multi-word queries, "princess diana" and "chat rooms".

From an efficiency point of view, the semantic relevance of the queries to the indexed collection is not terribly important. On the other hand, it is important that the queries have the right statistical properties. In particular, "inappropriate" queries can be processed at quite different throughput rates to "appropriate" ones, especially if "inappropriate" means "without many answers in the collection". Individual query terms in an inappropriate log may be much less frequent in the collection than those in an appropriate log, and there may be many fewer matching answers.

Fortunately, the wt10g TREC corpus does match the Excite97 query log, and can be used to gauge term statistics in the web as a whole. The wt10g collection is a 10 GB subset extracted from the 100 GB wt100g collection, with attention paid to ensuring coherence and document quality. The wt100g collection was crawled in 1997, and the documents were taken from the web as whole, not restricted to a particular set of domains. In terms of both date and domain wt10g is thus a good match for Excite97.

| Term | Collection | |
|---|---|---|
| | wt10g | GOV2 |
| "sex" | 1.87 | 1.39 |
| "free" | 13.59 | 6.70 |
| "nude" | 0.20 | 0.01 |
| "pictures" | 2.54 | 0.60 |
| "pics" | 0.23 | 0.02 |

*Table 1: Collection frequency $f_t$ as a percentage of the number of documents in the collection, for the TREC collections* wt10g *and* GOV2*, and the five most frequent terms in the* Excite97 *query log.*

Table 1 shows the five most frequent query terms in the Excite97 log and compares their occurrence frequencies in wt10g and GOV2. In all cases, the terms are less common in GOV2 than in wt10g, ranging from two-thirds to a twentieth of the frequency. This discrepancy means that, proportional to collection size, the Excite97 log should execute faster against the GOV2 collection than against the wt10g collection. To test this hypothesis, we extracted a slice of the GOV2 collection with the same number of documents as wt10g, and ran 10,000 Excite97 queries against it. The query stream took 16% longer against wt10g than against the GOV2 slice, confirming that the first set of experiments in which we applied the "real" Excite97 queries to the "real" GOV2 collection were probably biased.

In one sense, 16% is not that great a difference, and it could be argued that using the same queries in all runs is sufficient to guarantee fairness. However, Table 1 points to another characteristic of collection inappropriateness that was particularly relevant to our distributed experiments. Consider the notion of *workload*, the amount of work a term imposes upon the system during processing of a query set. Workload is the product of the term's frequency in the query set and its frequency in the collection, the latter measured concretely as the length (in bytes) of the term's inverted list in the index. That is, if $l(t)$ is the length of the term's inverted list in the index, and $f_q(t)$ is the number of times that term occurs in the query set, then the term's workload is given by $w(t) = l(t) \times f_q(t)$, and its proportional workload by

$$\frac{w(t)}{\sum_{t \in T_Q} w(t)}$$

where $T_Q$ is the vocabulary of the query stream. Since both collections and query sets have a skewed term frequency distribution, the workload of a query set's vocabulary should also be skewed.

The term "free" is the second most common term in the query log; occurs in almost a seventh of wt10g documents; and around half as frequently in GOV2. But for the 10,000 Excite97 queries being used in the experiments, "free" is the most workload intensive term, and generates 6.2% of the total workload for the wt10g collection, compared to 3.4% for GOV2.

The importance of workload skew in retrieval experiments should not be underestimated. Two of the three methods we were considering in our experiments involved partitioning the index between nodes by terms. Higher term workload skew means less evenness on average between partitions, which means in turn poorer distribution of processing workload between nodes. We eventually discovered that poor balancing of workload between nodes was the biggest problem in the pipelined system [Moffat et al., 2005].

Table 2 examines the issue of workload balancing in a term-partitioned system (such as pipelining). To create the table, the same sequence of 10,000 queries was extracted from the Excite97 log. The query term vo-

| Collection | Query set | 4 | 8 | 16 |
|---|---|---|---|---|
| wt10g | Excite97 | 1.209 | 1.455 | 1.943 |
| GOV2 | Excite97 | 1.171 | 1.356 | 1.684 |
| GOV2 | synq | 1.205 | 1.452 | 1.941 |

*Table 2: The ratio of maximum partition workload to the mean partition workload for different numbers of partitions, using* $10{,}000$ *random vocabulary partitionings, and* $10{,}000$ *queries.*

cabulary defined by those $10{,}000$ queries was then randomly partitioned across four, eight, and sixteen processors. The total workload at each node during a run of the $10{,}000$ queries was measured, first against wt10g, then against GOV2 (the third row is explained later). Finally, from this data the ratio of the most heavily loaded node to the mean of all nodes was computed. The figures given in Table 2 are the averages of these ratios over a set of $10{,}000$ different random term partitionings. For example, with $k = 4$ processors, on average a random vocabulary partitioning of wt10g resulted in one of the processors having an assigned workload (in a term-distributed sense) that was nearly $21\%$ greater than the average across the four processors.

In a term-partitioned system the most heavily loaded node is likely to become a bottleneck. Table 2 demonstrates, first, that the problem of workload imbalance grows as the collection is split into more parts; and second, that the imbalance grows more quickly when the query set is appropriate for the collection than when it is not. Use of an inappropriate query stream in our first set of experiments led us to erroneous conclusions about the scalability of the pipelining regime we were testing.

*Moral:* Mixing real-world data with inappropriate real-world queries may yield artificial outcomes.

## 5 Synthetic queries

Because GOV2 was the only large collection available, we had no choice but to persist in using it. In the absence of a matching query set – the log from a US government search engine would have been very useful – we decided to generate a synthetic query set that was statistically "appropriate" to GOV2.

Randomly generating queries is a fraught exercise. Consider, for example, a method that uniformly selects terms from the available vocabulary. Such a method would produce only incidental skew in the query term frequency distribution, and result in relatively balanced workload across the partitions. In particular, the median frequency of a term in the collection is small compared to the median collection-frequency of terms in typical query sets.

Synthetic query generation is nothing new in distributed information retrieval research. However, previous methods have been based on one of two unsatisfactory frequency distribution models. The first model is exactly the uniform distribution considered in the previous paragraph [Badue et al., 2001, Tomasic and García-Molina, 1993, Jeong and Omiecinski, 1995]. The second model incorporates skew; however, the skew is usually based upon the frequency of terms in the *collection*, either by directly following that distribution [Cahoon et al., 2000], or by ranking according to collection frequency then fitting to a Zipf distribution [Jeong and Omiecinski, 1995]. However, this is also unsatisfactory, since the correlation between term frequency in natural query logs and term frequency in document collections is relatively weak [Baeza-Yates, 2005]. More generally, there is a problem with attempting to fit query term frequency distributions to models like Zipf's law, even if the ranking and coefficient are based upon natural query logs. It is well understood that, though such power-law models may fit the bulk of the distribution quite well, they often fail to fit its extremes [Baeza-Yates, 2005].

Other work on distributed text query evaluation has used queries created by hand as part of standard query sets used for retrieval effectiveness tasks, such as those published by TREC [Ribeiro-Neto and Barbosa, 1998, Badue et al., 2001]. However, such query sets are very short, typically being composed of only 50 queries – hardly enough to allow the system to even get warmed up. Also, having been hand-crafted by people experienced in information retrieval, they tend to employ a more discriminating (and therefore lower-frequency) vocabulary than do natural query logs.

After considering these alternatives, we set four desiderata, in decreasing order of importance: (1) the query term frequency distribution should be appropriate for the collection; (2) the collection term frequency distribution should be appropriate; (3) the query length distribution should be appropriate; and (4) the query term co-occurrence distribution should be appropriate. Query coherence (or meaning) was not an important requirement, and because we were working with bag-of-word queries (and not phrase queries), it mattered little if the generated queries were nonsensical to human readers.

To meet the four requirements, a query translation method was developed based on term frequency. Real-world queries from the Excite97 log were translated on a term-by-term basis, by finding target terms in the target collection with similar frequencies to those of the source terms in the wt10g collection.

Supposing that $C'$ is a target collection for which queries are required, the details are as follows. The required inputs are an existing query set $Q$ and a source collection $C$ for which $Q$ is appropriate. As before, $T_Q$ is the set of terms occurring in $Q$; and $|C|$ and $|C'|$ are defined to be the number of documents in $C$ and $C'$. Similarly, $T_{C'}$ is taken to be the set of terms occurring in $C'$. For each $t \in T_Q$, a translation term $t' \in T_{C'}$ is picked such that $f_t/|C| \approx f_{t'}/|C'|$, where $f_{t'}$ is the term frequency in $C'$. This simple process maintains

| | |
|---|---|
| "spice sex" | "contra vhs" |
| "cartoon art" | "proposition claims" |
| "star trek" | "especially eliminated" |

*Figure 1: Sample query translations, converting* Excite97 *queries applied to* wt10g *(on the left) to synthetic equivalents applied to* GOV2*. For example, the term "*sex*", which occurs in* 1.87% *of* wt10g *documents but only* 1.39% *of* GOV2 *documents, is translated to "*vhs*", which occurs in* 1.78% *of* GOV2 *documents.*

identical query length and query term frequency distributions between $Q$ and $Q'$.

To additionally preserve workload characteristics and term co-occurrence rates, the translation process was performed one query at a time. For each $q \in Q$, there were some terms $t_1 \ldots t_i$ that already had translations $t'_1, \ldots, t'_i$, and others $t_{i+1} \ldots t_x$ that did not. To find a binding for $t_{i+1}$, a sequence of possible matching terms with the right collection frequency was explored. If the conjunctive Boolean query $\bigwedge_{j=1}^{x} t_j$ had no matches in $C$, then any suitable set of bindings was assigned to $t_{i+1} \ldots t_x$. On the other hand, if $\bigwedge_{j=1}^{x} t_j$ had a non-empty answer set in $C$, then at least three and as many as seven possible bindings for $t'_{i+1}$ were explored, and the one with the most answers to $\bigwedge_{j=1}^{i+1} t'_j$ in $C'$ chosen. This continued until either all terms in the query had mappings, or until no non-empty target query was uncovered after seven attempts to find a term binding, at which point the remaining terms were assigned any mapping of the right collection frequency.

The synthetic query set $Q'$ generated by this process is referred to as synq. Figure 1 shows three of the queries in synq and their original forms, and illustrates the fact that the translated queries generally do not make semantic sense. However, in conjunction with GOV2 they do match the source query set as it applies to wt10g in all of the statistical aspects that we are concerned with for our experiments. The last row of Table 2 measures the term-partitioned workload skew for synq on GOV2, and exhibits the same pattern of values as does the row that measures skew for the Excite97 queries on wt10g.

## 6 Careful scaling

Once the data and queries had been fixed, the next issue we grappled with was how exactly to structure experiments so that we correctly isolated scale as a factor – one of the hypotheses we sought to test was that pipelining was "more scalable" than document distribution.

In a uni-processing environment, scalability is established by working with increasing amounts of data, and measuring query throughput (or some related attribute such as average elapsed time). A strategy can claim to "scale well" if, when normalized by the data volume, the throughput rate stays steady or decreases. For example, if a throughput of $X$ queries per second is possible on $G$ gigabytes of data, then a scalable system should deliver $X/10$ queries per second (or more) on $10G$ GB of data.

In a distributed experiment, it is tempting to apply scale incorrectly, and to seek to verify that if one machine can attain $X$ queries per second on $G$ GB of data, then $k$ machines can attain $kX$ queries per second on $G$ GB. In fact, the correct experiment is to apply $k$ machines to $kG$ GB and establish that the rate of $X$ queries per second can be maintained in the face of data growth. At face value these are the same experiment, but there is a subtle difference between the two. When the number of processors changes, and the volume of data is held fixed, other effects can intrude. For example, the $k$ machines also have $k$ times as much memory, meaning that a greater fraction of the index can be in memory.

Indeed, if $k$ machines cannot process $kG$ GB at $X$ queries per second, and one machine can process $kG$ GB at $X/k$ queries per second, then distribution is a failure, since a $k$-way mirrored system is superior.

In our final set of experiments with GOV2 we were careful to work with homogeneous fractions of the collection, and measured query throughput rates using one processor and $1/8$th of the collection, two processors and $1/4$ of the collection, four processors and $1/2$ of the collection, and eight processors with all of the collection, so as to correctly identify the effects of scale.

*Moral:* When testing that a process is scalable, be sure that you know what you are actually measuring.

The scaling strategy described in the previous section led to another uniformity issue to do with the mechanics of the query engine. The Zettair system uses a dynamic thresholding scheme that limits the number of accumulators used during processing, in order to bound the amount of query-time memory needed [Moffat and Zobel, 1996, Lester et al., 2005]. Limiting memory use is important if throughput is to be maximized, since many concurrent query threads are likely to be active at any given time. In preliminary experiments with a monolithic system we had found that a limit of $L = 100{,}000$ accumulators was sufficient with GOV2 to achieve a high level of retrieval effectiveness (measured using average precision over a fixed query set, relative to an unrestricted run).

Initially, we applied this limit to each processing node. However, this was unfair to the document-partitioned system. It was forced to maintain $kL$ accumulators system-wide during its (parallelized) processing of each query, compared to $L$ for (serialized) pipelining. That is, we were placing the baseline system at an unfair disadvantage. To be fair to the document-partitioned architecture, the per-node accumulator limit should be set at $L/k$, which then raises the question as to how retrieval effectiveness behaves. Unfortunately, by this stage we were using synthetic queries, and were unable to assess retrieval effectiveness. Instead, we quantified the extent to

which the rankings varied from each other using a dissimilarity metric. Additional experiments then justified the correctness of the scaled-accumulator approach; and a $k$-processor document-partitioned split of GOV2 with a $100{,}000/k$ per-node accumulator limit yielded almost exactly the same rankings as a monolithic system with a $100{,}000$ limit.

The second way in which accumulator limits might be scaled is with the size of the collection. When the collection size halves, the target $L$ might also be halved. Again, it came down to the effectiveness results; and in this case, they were more equivocal. The answer rankings (to depth $r = 1{,}000$) did change if the accumulator limit was decreased in proportion to the size of the collection; however, the difference was slight (by the normalized dissimilarity measure, roughly $6\%$ for the smallest collection). This relatively small decrease in effectiveness was acceptable, given the need for consistency in the experiments.

*Moral:* If in doubt, make choices in a manner that least disadvantages the attribute of the baseline system against which you most wish to compare.

## 7 Disk placement

As the various inputs to the experiments were refined, variability in the timing results became apparent. The largest index was $16$ GB, and the cluster machines each had $1$ GB of main memory. Overall performance was thus dependent on the performance of disk reads, and the variability we were observing seemed in some way connected with disk attributes.

Disk performance is primarily affected by two factors: the degree to which the stored data is fragmented into groups of blocks or "extents"; and the physical location on disk of those extents. Fragmentation causes delays as the disk head seeks from one extent to another; the greater the distance between the fragments, the greater the delay. And disk platters rotate at a constant speed, so a disk head takes the same amount of time to complete a rotation at the rim of the platter as it does at its spindle; but more blocks are held per circumference at the rim than at the spindle, due to the greater radial area. It turned out that on the disks we were using, the read-speed ratio between rim-near and spindle-near blocks was around $7{:}4$.

Our initial experiments had been done on machines with a single $220$ GB partition for experimental data that was shared with other active users. In particular, it was not practicable to keep it empty of data for the many months of our experiments. The large partition size made our experiments subject to wide variability in disk read speeds between runs; and the presence of other data on disk made file fragmentation more likely. As a result, our initial results suffered from significant I/O-related variability. For example, the system running the full GOV2 index on a single machine could spend anywhere between $5\%$ and $25\%$ of its time waiting for disk. This $20\%$ difference in effective processing capacity led in turn to a similar difference in system throughput, masking the effects we were trying to measure. The I/O variance is especially problematic in a situation such as ours, where the measurements of necessity involve different files of different sizes.

Our first approach to solving this issue was to pre-allocate "storage areas" by creating a number of $4$ GB files filled with null bytes on every machine, then reusing these files for each experiment by writing the index data into their existing blocks. In UNIX terms, this means opening the files for writing without setting the `O_TRUNC` flag. (An `rcp` replacement that behaved in this way was developed.) Each time a given index was loaded onto a server, it was then guaranteed to be in the same location. It was not a sufficient solution in itself, however, and did not guarantee that indexes of different sizes received equivalent placement. As an extreme, imagine a file where the first $2$ GB were contiguous and rim-wards, but the remaining $2$ GB were fragmented across small extents close to the spindle: a $2$ GB index and a $4$ GB index written into such a storage space would have very different read performance. In addition, file pre-allocation also does not guarantee that the storage spaces on different machines are equivalent.

Storage pre-allocation and reuse, therefore, is not in itself sufficient; that storage space must also be similarly located on different systems, and substantially contiguous in each case. Actually getting equivalent file allocations across different machines, particularly on very large, partially-full partitions, then became a hit-and-miss process of iteratively creating files, finding ones with similar-looking block locations and reasonable contiguity, and then verifying their equivalence by timing tests – a rather tedious and time-consuming process.

*Moral:* Make sure that the baseline system and the system being compared against it are given equal access to resources.

Having battled to create appropriate storage spaces, we decided that such a process could hardly represent either experimental best-practice or a practical example to others, and in the end we adopted another solution.

The final method we used to control disk location variability was to re-partition all of the disks, and on each machine create a partition just a little larger than the largest index. Constraining the size of the experimental partition meant that the range of possible file locations, and therefore disk read speeds, was similarly constrained. We placed this partition in the fastest section of available disk space, which is also the section of disk with the most blocks per circumference, and hence the smallest speed difference between contiguous blocks. With a dedicated experimental partition, it was then possible to always copy indexes into an initially empty partition.

*Moral:* If performance is to be repeatable, disk partitions should be kept as small as possible, and should be empty at the start of each run.

Even with all this preparation, we found that XFS, the filesystem we were using, had some peccadilloes that surprised. The first was that even on an empty disk, it does not always write files contiguously, but sometimes leaves moderately sized gaps. (Block allocations under XFS filesystem can be retrieved with the `xfs_bmap(8)` utility.) We therefore automated our index installation tool to check the block allocations, and recopy the files if they contained a gap beyond a certain tolerance (we used $0.5$ GB). This step could be omitted by experimenters less embittered about the whole issue than we were by this stage. As for most modern filesystems, there is no filesystem editor available for the XFS filesystem, nor is there any way to force particular block allocations for a given file.

A second peculiarity of XFS's behavior was that it regarded the first block of the partition to be adjacent to the last, and if it started a large file towards the end of a partition, would "contiguously" place the rest of the file at the start of the partition. Moreover, if files and directories were iteratively created and deleted, (which was our initial procedure with indexes), each new directory's files were placed closer to the spindle than the previous directories files, until the end of the partition was reached and the addresses wrapped around. The result was that, over a set of experiments, a sequence of more-or-less contiguous indexes was created, interspersed at regular intervals with a single maximally-gapped one. We discovered this behavior via puzzling cyclical breakdowns in performance over sets of runs.

To counterbalance these annoyances, XFS does have a more useful trait: if all the files in a directory are deleted, but not the directory itself, then any new files created in that directory are located starting at the same block offset as the old ones. We exploited this behavior by creating a couple of dozen directories on each node, and in each directory, creating a file holding a single byte. The directory with the file in the rim-closest block was retained; the remainder were deleted. Then, for all of the experiments, index files were copied into that directory, and accessed via symlinks.

*Moral:* You may need to become intimately acquainted with the behavior of your disk drive.

During the sequence of extensive timings we discovered one more problem that threatened our sanity – the disk on one of the nodes was $4.5\%$ slower than the disks on the other seven. This was despite the fact that all nodes were purchased at the same time in a single order and with supposedly identical configurations. We contemplated deliberately placing the data partition on this disk in a more rim-ward location than on the other nodes, but in the end decided that life was too short, and instead avoided using this machine for all except the less disk-intensive $k = 8$-node runs.

*Moral:* There is always something else to go wrong.

## 8 Dynamic thresholding

There were also useful outputs from our tribulations.

The most tangible of these related to the horrible realization that led to the withdrawal of our second submission of the pipelining paper. The issue in question was algorithmic in nature, and concerned how `Zettair` was implemented. The solution to that problem is now the subject of a separate paper [Lester et al., 2005]; this section briefly summarizes the basis of that work.

Dynamic thresholding was mentioned in Section 3 as a technique for limiting memory usage. The `Zettair` system nominally uses the `continue` accumulator strategy [Moffat and Zobel, 1996]. Under this scheme, postings are permitted to create new accumulators until the limit is reached; thereafter postings may update the scores of existing accumulators, but not cause the creation of new ones. The original design of these algorithms was that the transition between the initial "OR" state and the final "AND" state be made at the end of processing an inverted list. However, the implementation in `Zettair` departed from this by making the transition as soon as the accumulator limit $L$ was hit, even if this was in the middle of an inverted list – a seemingly innocent "interpretation" of the policy that had the benefit of making $L$ a firm limit, rather than a target that might be grossly exceeded.

However, changing states in the middle of inverted lists has the undesirable effect of favoring documents that appear early in the collection, and has an adverse impact on retrieval effectiveness. More critical from our point of view was the consequent behavior of `Zettair` – it contained an optimization that aborted processing an inverted list when the largest document number in an accumulator had been surpassed within that list. In all of the systems, if the first term in a query was the one that triggered the state change, then the largest document number in the accumulator could only be part-way through the collection, and the `Zettair` optimization meant that only the first part of each other inverted list was dealt with.

Even worse, in the pipelined system, the variable recording "largest current document number in the accumulator set" was being passed from machine to machine as part of the query bundle without being properly initialized, and once the accumulator limit had been reached, the inverted lists for subsequent terms were being fetched from disk, but not processed in any way. That is, the system was reverting to the `quit` strategy of Moffat and Zobel, magnifying apparent throughput and giving pipelining an absolute advantage. It was this complete breakdown in experimental rigor that caused us to withdraw the second submission of our paper.

The chain of events involved in this error was uncovered only when, post submission, we thought to check again that the various schemes were identifying roughly the same set of documents, and found that

they weren't. The moral of this section is one that even beginning software engineers are taught:

> *Moral:* After making a change to a program, rerun all of the tests, not just the most recent one.

## 9 Parallelization

One area that we did get right was that of parallelized query processing. Much of the work in the area of text query processing efficiency has employed a serial processing model, with each query required to complete before the next one is commenced. (Indeed, this was a requirement of the 2005 TREC Terabyte efficiency track). In this model, the objective is to minimize average query response time.

Our investigation allowed the parallel processing of queries, and concentrated instead on minimizing the elapsed time to process a whole batch of queries. In a parallelized environment, average query response time and query throughput rate are not the same. Indeed, one generally maximizes throughput by allowing a high degree of parallelism, at the expense of average query response time. In a sense, the decision to support parallelism was an easy one to make – right from the planning stage, it was clear that pipelining would not come close to matching document-distribution in a serial-evaluation environment.

All three distribution methods were treated equally. The public `Zettair` software does not support parallelism, so modifications had to be added to enable multi-threading. Experiments were then run against each architecture with different numbers of simultaneously active queries, to find the number that maximized throughput, which turned out to be 32 simultaneous queries for all methods. Parallelism allowed both document-partitioning and pipelining to greatly increase throughput compared to serial mode evaluation; using eight nodes, the former increased throughput by over 150%, the latter by almost 600%. Even a monolithic `Zettair` system running on a (hyper-threaded) single-processor system gained a 70% increase in throughput from parallelization.

> *Moral:* Even on single-processor machines, it is unlikely that maximum performance can be attained without parallelization.

## 10 Conclusion

The initial experiments were promising for the new pipelined architecture, but as we refined our experimental design and techniques, the promise slowly evaporated. We did, however, develop a much richer understanding of how these experiments should be performed, and a much keener appreciation of the important issues in distributed text query evaluation. We now have a strong basis for further research into different techniques for index partitioning and workload balancing in distributed retrieval environments.

**Acknowledgments** This work was supported by the Australian Research Council. Justin Zobel (RMIT University) provided helpful input.

## References

C. Badue, R. Baeza-Yates, B. Ribeiro-Neto, and N. Ziviani. Distributed query processing using partitioned inverted files. In G. Navarro, editor, *Proc. Symp. String Processing and Information Retrieval*, pages 10–20, Laguna de San Rafael, Chile, November 2001.

R. Baeza-Yates. Web usage mining in search engines. In A. Scime, editor, *Web Minng: Applications and Techniques*. Idea Group Publishing, 2005.

B. Cahoon, K. S. McKinley, and Z. Lu. Evaluating the performance of distributed architectures for information retrieval using a variety of workloads. *ACM Transactions on Information Systems*, 18(1): 1–43, January 2000.

B. P. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *ACM SIGIR Forum*, 32(1):5–17, Spring 1998.

B.-S. Jeong and E. Omiecinski. Inverted file partitioning schemes in multiple disk systems. *IEEE Transactions on Parallel and Distributed Systems*, 6(2):142–153, 1995.

N. Lester, A. Moffat, W. Webber, and J. Zobel. Space-limited ranked query evaluation using adaptive pruning. In *Proc. 6th Int. Conf. on Web Informations Systems*, pages 470–477, New York, November 2005. LNCS 3806, Springer.

A. Moffat, W. Webber, J. Zobel, and R. Baeza-Yates. A pipelined architecture for distributed text query evaluation. September 2005. Submitted.

A. Moffat and J. Zobel. Self-indexing inverted files for fast text retrieval. *ACM Transactions on Information Systems*, 14(4):349–379, October 1996.

A. Moffat and J. Zobel. What does it mean to "measure performance"? In X. Zhou, S. Su, M. P. Papazoglou, M. E. Owlowska, and K. Jeffrey, editors, *Proc. 5th Int. Conf. on Web Informations Systems*, pages 1–12, Brisbane, Australia, November 2004. LNCS 3306, Springer.

B. A. Ribeiro-Neto and R. R. Barbosa. Query performance for tightly coupled distributed digital libraries. In *Proc. 3rd ACM Conference on Digital Libraries*, pages 182–190, Pittsburgh, PA, June 1998. ACM Press, New York.

A. Spink, D. Wolfram, B. J. Jansen, and T. Saracevic. Searching the web: The public and their queries. *Journal of the American Society for Information Science*, 52(3):226–234, 2001.

A. Tomasic and H. García-Molina. Performance of inverted indices in shared-nothing distributed text document information retrieval systems. In M. J. Carey and P. Valduriez, editors, *Proc. 2nd International Conference On Parallel and Distributed Information Systems*, pages 8–17, Los Alamitos, CA, January 1993. IEEE Computer Society Press.

J. Zobel, A. Moffat, and K. Ramamohanarao. Guidelines for presentation and comparison of indexing techniques. *ACM SIGMOD Record*, 25(3):10–15, October 1996.

# Document Expansion versus Query Expansion for Ad-hoc Retrieval

*Bodo Billerbeck* *Justin Zobel*

School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne, Australia
{bodob,jz}@cs.rmit.edu.au

November 18, 2005

**Abstract** *In document information retrieval, the terminology given by a user may not match the terminology of a relevant document. Query expansion seeks to address this mismatch; it can significantly increase effectiveness, but is slow and resource-intensive. We investigate the use of document expansion as an alternative, in which documents are augmented with related terms extracted from the corpus during indexing, and the overheads at query time are small. We propose and explore a range of corpus-based document expansion techniques and compare them to corpus-based query expansion on TREC data. These experiments show that document expansion delivers at best limited benefits, while query expansion – including standard techniques and efficient approaches described in recent work – delivers consistent gains. We conclude that document expansion is unpromising, but it is likely that the efficiency of query expansion can be further improved.*

**Keywords** Document expansion, automatic query expansion, pseudo relevance feedback, efficiency

## 1 Introduction

Word mismatch is a common problem in information retrieval. Most retrieval systems match documents and queries on a syntactic level, that is, the underlying assumption is that relevant documents contain exactly those terms that a user chooses for the query. However, a relevant document might not contain the query words as given by the user. Query expansion (QE) is intended to address this issue. Other topical terms are located in the corpus or an external resource and are appended to the original query, in the hope of finding documents that do not contain any of the query terms or of re-ranking documents that contain some query terms but have not scored highly.

A disadvantage of QE is the inherent inefficiency of reformulating a query. With the exception of our earlier work [2], these inefficiencies have largely not been investigated. In this work we proposed improvements to the efficiency of QE by keeping a brief summary of each document in the collection in memory, so that during the expansion process no time-consuming disk accesses need to be made. While some of the methods proposed in this earlier research more or less maintain effectiveness, the process is sped up by roughly two-thirds. However, expanding queries using the best of these methods still takes significantly longer than evaluating queries without expansion.

In this paper, we explore the use of document expansion (DE) as an alternative to QE. In DE, documents are enriched with related terms. Although, while not prohibitively so, there is a significant cost associated with expanding documents; this is undertaken at indexing time, and there is only marginal cost at retrieval time. In principle it is reasonable to suppose that DE will help resolve the problem of vocabulary mismatch and thus yield benefits like those obtainable with QE.

We propose two new corpus-based methods for DE. The first method is based on adding terms to documents in a process that is analogous to QE: each document is run as a query and is subsequently augmented by expansion terms. The second method is based on regarding each vocabulary term as a query, which is expanded and used to rank documents. The original query term is then added to the top-ranked documents.

Our experiments measure the efficiency and effectiveness of QE and DE on several collections and query sets. We find that, on balance, DE leads to improvements in effectiveness, but few of the measured gains are statistically significant; the computational cost at query time is small. In contrast, both standard QE and the efficient QE that we proposed earlier [2] lead to gains in most cases, many of them significant, while the efficient QE is less than twice the cost of querying without expansion.

Our experiments were, within the constraints of our resources, reasonably exhaustive. We tested several alternative configurations of DE and explored the parameters, but did not observe useful gains in effectiveness. We conclude that corpus-based DE is unpromising for small sets of terms. We did not explore QE to the same extent, yet found effectiveness to consistently improve, and thus believe that further gains in performance may be available.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.**

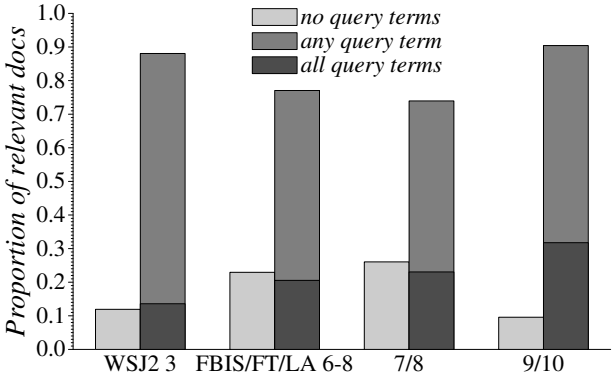



Figure 1: *The proportion of relevant documents that contain none, any, or all query terms over all title queries for each data set as shown (collections and queries are discussed in Section 4). Stopping, but no stemming, was used to generate this graph.*

## 2 Background

User queries often do not reflect the exact terminology of a document. Whereas a document might be on the exact topic of a query, this document will not be retrieved if it doesn't contain any of the key words in the user query. Figure 1 shows that as many as 25% of documents that are judged to be relevant do not contain *any* terms that appear in a concise query. The actual proportion might be much larger than this figure suggests, since in the TREC framework – from which the graph was produced – the relevance of only a relatively small number of documents is judged for each query. QE adds related terms to a query, so that those documents will be included in the ranking.

Early successful attempts of QE were based on relevance feedback [12]. Since this required the user to assess a large number of documents for their relevance, it proved to be impractical. Rather than asking users to assess whole documents, *interactive QE* suggests topical terms to the user that do not appear in their query. The user is then able to add any number of those terms to the query. Since users are generally reluctant to provide such information, and it was found that algorithms are just as likely as non-expert users to pick terms that enhance (or conversely, do not enhance) retrieval [13, 18], research has since shifted to *pseudo relevance feedback*. Terms, that are heuristically found to be related to the topic of the original query are automatically added to the query, without user intervention.

One approach to *automatic query expansion* methods – that require no user input other than the original query – is *global analysis* where collections are analysed using metrics such as term co-occurrences. Correlating terms are then used to build a thesaurus which is drawn on during query time by simply looking up related terms that are subsequently appended to the query.

*Local analysis* methods (such as that proposed by Robertson and Walker [11]) retrieve a set of documents through an initial ranking by the original query (see Algorithm 1 and Figure 2a). Terms from those documents

**Algorithm 1** Conventional QE through local analysis

1: run original query $q$ and rank docs in collection
2: select top 10 documents as local set $R$
3: extracted all terms $t$ from local set $R$
4: **for all** terms $t \in R$ **do**
5: calculate term selection value
6: **end for**
7: rank terms $t$ based on their a term selection value
8: add top $|E|$ terms to the original query
9: run expanded query $q'$ and rank documents

are retrieved. The frequency of those terms amongst the set of retrieved documents as well as collection statistics are taken into account in order to determine which terms are added to the query.

While global analysis mechanisms are inherently much more efficient than local ones (only dictionary lookups are performed during query time, rather than costly document retrieval and parsing), they are also likely to be less successful [19]. The difference in effectiveness is based on the problem that a term can take on different meanings, depending on which context it appears. Local analysis methods inherently disambiguate word senses better, as expansion terms are sourced from documents that are retrieved with the whole query, rather than individual query terms. Because of this difference in performance and the fact that our methods proposed below are based on local analysis blind relevance feedback, we compare the effectiveness and efficiency of our proposed DE techniques to that of a standard local analysis technique.

## Improving QE efficiency

Local analysis QE consists of several steps, some of which are time-consuming. First, there is an initial ranking process, where documents are identified that are presumably on the topic of the query. Next those documents are retrieved. Since most queries will rank different documents, these documents are most likely not cached (assuming a reasonable amount of memory) and have to be fetched from disk at a significant penalty in time. This is the most costly subtask of the QE process. Once documents are in main memory, they have to be parsed and statistics of term occurrences in respect of the local set of documents have to be computed. At a relatively minor cost, statistics of those terms for the whole document collection have to be looked up. Terms are then chosen and appended to the query. Finally the query has to be re-run, which requires not only the re-processing of inverted lists for the original query terms, but new lists have to be retrieved, decoded, and analysed.

Only the first step needs to be performed in the absence of expansion. There is no previous research concerned with accelerating the QE process in information retrieval, apart from our earlier paper [2], where we use a summary of each document consisting of that

document's top *tf.idf* terms. During querying, a fixed number of terms – or alternatively, terms with a *tf.idf* value above a certain threshold – is kept in memory for each document. While performing local analysis, rather than retrieving documents from disk, the in-memory summaries are referenced. This procedure improves querying throughput by a factor of two, while effectiveness is only marginally degraded. Although they were able to avoid the time-consuming retrieval of documents from disk, they restricted their focus to standard approaches to QE.

## Document expansion

Whereas DE has recently been applied in various areas of information retrieval, it has not been used instead of QE to improve ranking effectiveness, with the exception of Ide and Salton [5]. While not actually expanding documents, Ide and Salton manipulate their vector representation not unlike the DE methods proposed in this paper, although – unlike in this paper – they use actual relevance feedback. They propose to change the document vector space so that relevant documents are closer to the query vector. They achieve improvements of 10% to 15%.

Actual DE (that is, not just manipulating document vectors, but actually adding terms to documents) was first used by Singhal and Pereira [15] in the context of speech retrieval. Since speech recognition is unreliable (at the time of publishing, Singhal and Pereira report error rates of up to 60% for particular collections – although speech recognition has improved since), transcribed documents are expanded with related terms from a side corpus. Singhal and Pereira achieve a relative increase in average precision of 12% in addition to employing pseudo relevance feedback based on the technique proposed by Rocchio [12].

*Latent semantic indexing* [4] is in effect a DE method, however for information retrieval it was found to be inferior to the vector space model [9].

Li and Meng [8] use DE for spoken document retrieval with good improvements in Cantonese monolingual retrieval and in Mandarin cross-language retrieval.

Both Lester and Williams [6] and Levow and Oard [7] have used DE for topic tracking. Whereas Lester and Williams use DE to enrich topic profiles and do not specify whether it bears any benefit, the latter get consistent improvements in Mandarin cross-lingual retrieval by expanding the documents to be tracked.

With the exception of Lester and Williams (who expand only translated documents), all other work mentioned above uses DE in the context of enriching possibly incorrectly translated documents.

## Query associations

One of our proposed DE methods (detailed in Section 3) is in essence quite similar to *query association* as used in the context of effective retrieval [14]. We describe these here and highlight the differences to our proposed method later. Scholer et al. make use of a query log, by running each query of the log and adding the text of the query to the top $N$ ranked documents. Each document is augmented with the top $M$ queries that achieved the highest similarity score. They found that good values for $M$ and $N$ are 19 and 39 respectively.

We previously made use of query associations in conjunction with QE [1] with good success, however, for that work we stored associations separately and then expanded queries from the especially created *surrogates* conventionally.

# 3 Document expansion methods

Rather than expanding a query from an initially retrieved set of documents, which is time-consuming, DE expands documents with potential query terms that occur in similar documents. While this expansion process is reasonably costly, it is done prior to indexing time. Query times are only slightly increased, since inverted lists are on average, say 10% longer, depending on which DE method is chosen.

There are several ways to expand documents. All methods have one aim: to eliminate inefficient run-time QE, while getting some effectiveness of a local analysis mechanism. Each of the following proposed methods makes use of local analysis at indexing time and expands the original documents with additional terms.

## Selection and weighting measures

Before describing the different DE techniques we propose, we first explain underlying equations needed to arrive at expanded documents. We use one similarity measure, three different measures to select expansion terms, and one measure that weights selected terms.

**Similarity measure.** To measure the similarity of queries to documents, we use Okapi BM25 [16] in all our experiments, where constants $k_1$ and $b$ are set to $1.2$ and $0.75$ respectively. We set $k_3$ to 0, motivated by the assumption that each term in contemporary queries [17] only occurs once.

**Term selection measures.** Depending on the expansion method, we use different measures to select terms from a set of candidate terms.

We use the *term selection value* [11] in our experiments for ranking terms, if not stated otherwise:

$$TSV_t = \left(\frac{f_t}{N}\right)^{f_{r,t}} \binom{|R|}{f_{r,t}}$$

where $f_t$ is the number of documents in the collection in which term $t$ occurs in, $N$ is the total number of documents in the collection, and $f_{r,t}$ is the number of the $|R|$ top ranked documents in which term $t$ occurs.

An alternative is the *Kullback-Leibler divergence*, which specifies the distance between two probability densities. In other words, each term in the local set of documents ($R$) gets a value associated with the relative rareness of a term in the current set as opposed to the whole collection. The *KLD* weight of terms that occur

**Algorithm 2** Document centric expansion

1: **for all** documents $d \in$ collection **do**
2: formulate query $q$ to consist of all terms $t$ in $d$
3: rank documents in collection against $q$
4: select top 10 docs (other than $d$) as local set $R$
5: using *TSV*, select top $|E| = 25$ terms from $R$ (excluding $t \in q \cap d$) and append to $d$
6: **end for**

relatively often (or seldom) in the local set in contrast to the entire collection will receive a higher (lower) value than terms that appear as often as their term frequency suggests. The *KLD* can be calculated as [3, page 154]:

$$KLD_t = \frac{f_{r,t}}{|R|} \times \log\left(\frac{f_{r,t}}{|R|} \times \frac{F + 0.01 \times |V|}{F_t + 0.01}\right)$$

where $F_t$ is the total number of occurrences of term $t$ in the collection, $F$ is the combined total number of occurrences of all terms in the collection, and $|V|$ is the number of unique terms in the collection.

**Term weighting.** In all our experiments, expansion terms are weighted[1] by the Robertson/Sparck Jones relevance weight [10], to be used in the Okapi formula:

$$rw_t = \frac{1}{3}\log\frac{(f_{r,t} + 0.5)(N - f_t - |R| + f_{r,t} + 0.5)}{(|R| - f_{r,t} + 0.5)(f_t - f_{r,t} + 0.5)}$$

## Document centric DE

For this DE technique each complete document is run as a query and the top $|E|$ expansion terms determined through local analysis are appended to the document (see also Algorithm 2 and Figure 2b). This method is conceptually similar to conventional QE. Although this way of expanding documents is reasonably time consuming, it could be sped up considerably by for instance using only the top $n$ *tf.idf* terms for each query.

Even though the Okapi variant that we use for our experiments is not well suited for queries with duplicate query terms, we found that using the standard BM25 formulation or the Cosine measure degrades results considerably. Using our training data, we found that selecting terms with the *KLD* worked consistently better than using their *tf.idf* value or *TSV*. Interestingly, it became clear that allowing terms which are already in a document to be appended to this document decreases effectiveness compared to restricting additions to new material. We also found that augmenting a document with 10% of the number of tokens in a document works best, rather than adding a fixed number of terms or using a global threshold value for the selection value of each candidate term. That is, a document that contains 100 words is augmented with 10 more words. A side effect of DE is therefore that document collections and associated indexes are roughly 10% longer than the original collection and indexes after expansion.

A potential problem with DE is that terms that are used for augmenting documents tend to be quite rare across the collection. Adding rare terms to documents means that, after expansion, those terms will be less rare, which will have an effect on retrieval performance.

[1] The dampening factor of 1/3 helps to prevent query drift. It was recommended by unpublished correspondence with the authors.

**Algorithm 3** DE based on vocabulary

1: **for all** words $t \in$ vocabulary $V$ **do**
2: form query $q$ from $t$
3: rank documents $d$ against $q$
4: select top 10 documents $R$ as local set
5: rank candidate terms using *TSV*
6: append top $|E| = 25$ terms to $q$, forming $q'$
7: rank 100 documents ($X$) against $q'$
8: **for all** documents $d$ in $X$ **do**
9: calculate $s$, the similarity score of $q'$ to $d$
10: save $t, d, s$ triplets
11: **end for**
12: **end for**
13: **for all** documents $d$ in collection **do**
14: select $0.1 \times |d|$ terms with highest $s$ and add to $d$
15: **end for**

## Term centric DE

This approach to document expansion mimics more closely a reversal of the conventional local analysis algorithm. Imagine a query that consists of one term only. The role of QE is to identify documents that are *about* this term, but do not necessarily include this term. This is done by adding terms to the query that co-occur with the query term within the local set. After expansion we therefore retrieve documents that do not contain the query term but that do contain expansion terms. DE inverts this scenario: it puts the query term into those documents that contain the expansion terms. This has the effect of adding potential query terms that are on the same topic as the document but are missing from it. In other words, Algorithm 3 ideally adds terms to a document that would have lead to the document being ranked if the term had been run as a single original term in an expanded query (see also Figure 2c). Our hypothesis is that this algorithm is a good match for queries consisting of single terms, however less so for the case of multi-term queries.

This expansion method is considerably faster than the document centric approach. However, a problem that does not occur with the document centric approach arises in a setting where the collection grows, such as the web. Since the basis for selecting expansion terms is changing with the addition of new documents, the terms previously chosen for a particular document might be sub-optimal. Furthermore, it is difficult to determine the best expansion terms for added documents, as those documents did not exist when the collection was originally ranked against terms. An – admittedly expansive – solution is to rerun the DE process after a certain number of documents have been added. Possible optimisations are outside the scope of this paper.

In our experiments, we rank 100 documents against an expanded term, although we found that ranking any

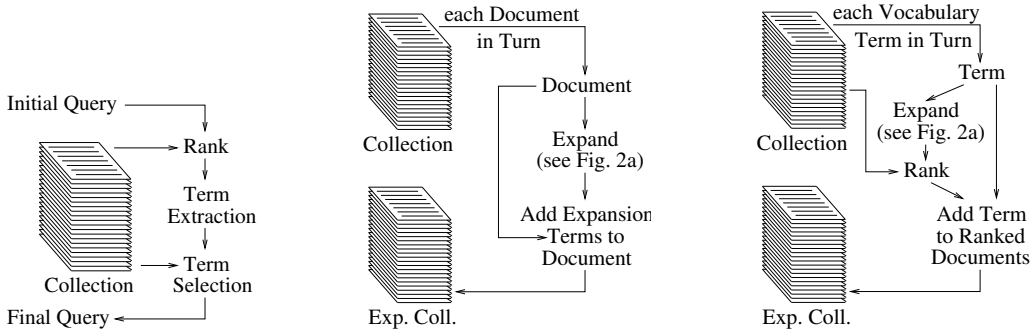



Figure 2: *The figure on the left (a) shows the central part of any expansion process proposed in this paper. The centre figure (b) shows how documents can be expanded by running each document as a query and adding the expansion terms back into the document. On the right (c) is shown how documents are expanded by running vocabulary terms as a queries and adding the terms to the top ranked documents.*

number between 90 and 110 documents works equally well. Contrary to the first method detailed above, we found that using the *TSV* to select terms worked better than choosing candidate terms based on their *KLD* values. We allowed terms to be added to documents even though they might already appear in that document. Surprisingly, we found that excluding terms on this basis leads to lower increases in effectiveness.

Query associations, as described in Section 2, differ from our DE methods in several ways. First, our techniques have the extra step of expanding a query with terms before ranking documents that a query gets associated with. Second, associations are based on external information in the form of query logs, whereas DE relies on within-collection data and statistics only. More importantly, the query association results [14] show that augmenting documents with queries works best when placing the restriction on queries to be associated with a document that all query terms must be present in the document. The effect of this restriction is that pertinent terms in a document get emphasised (their term count is increased and therefore the ranking of those documents is improved subsequently), rather than new terms – which address the problem of vocabulary mismatch – are added to the document. Adding new terms makes a document retrievable to queries that originally would not have ranked this document, even though it may be on the same topic.

## Expansion via phrases as queries

As an extension to the term centric expansion, instead of running individual vocabulary terms against the corpus to establish associations between those terms and documents, phrases can be used. This addresses a potential shortfall of the method above, which is a good match for queries consisting of single terms only.

We consider a phrase to consist of two or more contiguous terms that are not separated by either a stop word, an HTML or TREC tag, or any of the following characters: ?!,;:(){}[]. In separate experiments, we use maximal-length phrases and overlapping two-term phrases. As in the previous method, the phrases then get added to documents.

# 4 Experimental setup

We evaluate the proposed approaches in respect of effectiveness and efficiency as well as significance of results. As the underlying search engine we use Zettair.[2] We did not use stemming, but stopped queries. Although the local analysis parameters $|E|$ and $|R|$ are collection dependent, we did not tune those for each collection. Instead we use the default parameters of 25 and 10, respectively, in all cases.

**Test collections.** All our test data is drawn from the TREC conferences. To tune parameters and choose selection measures we used the Wall Street Journal articles from TREC disk 2, which covers issues from years 1990-92, referred to as WSJ2. With this collection we used the TREC 3 topics 151–200. We ran the title field as queries in all experiments to evaluate our system.

We used several collections to evaluate our techniques. One is sourced from the same TREC: Associated Press (AP); we used the AP data from disks 1 and 2 to match the TREC 3 topics and relevance judgements. We also used the newswire collection from TRECs 7 and 8 (NW). This collection is drawn from disks 4 and 5, without the congressional record. NW was used as a whole and also as several sub-collections from this collection, namely the Financial Times 1991-94 (FT), the Foreign Broadcast Information Service (FBIS) and the LA Times (LA). Testing was done against topic sets of TRECs 6, 7 and 8.

**Timings.** For timings, we used 100,000 stopped queries taken from two query logs collected for the Excite search engine [17]. Although these queries are web queries and not ideally suited to match the newswire data (we were not able to obtain a more suitable query log), these queries are adequate for testing the throughput only – rather than effectiveness – of the system.

Our timings were produced on two machines. The first is a Pentium IV 2.8 GHz machine with hyper-threading and 2 GB of main memory. The second is a dual Pentium III 866 MHz with 768 MB of main memory. In Table 1 these are denoted as *Lrg*

[2]Zettair is an open source search engine available from `http://www.seg.rmit.edu.au/`

| Expansion | WJS2 | | AP | | NW | | FBIS | | FT | | LA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Lrg | Ltl | Lrg | Ltl | Lrg | Ltl | Lrg | Ltl | Lrg | Ltl | Lrg | Ltl |
| None | 4.7 | 7.4 | 6.9 | 11.7 | 11.4 | 22.8 | 5.0 | 8.0 | 6.8 | 12.1 | 6.8 | 11.3 |
| QE | 25.4 | 47.1 | 29.0 | 52.5 | 145.9 | 211.2 | 49.4 | 123.5 | 41.2 | 87.6 | 32.6 | 62.3 |
| $S = 40$ | 7.5 | 14.8 | 11.4 | 22.5 | 20.9 | 52.0 | 8.1 | 16.3 | 11.8 | 24.4 | 10.6 | 20.6 |
| $Q = D$ | 5.3 | 8.6 | 7.7 | 13.3 | 12.1 | 24.8 | 5.5 | 9.4 | 7.6 | 13.7 | 7.3 | 13.3 |
| $Q \in V$ | 4.9 | 7.4 | 7.0 | 11.8 | 11.7 | 22.9 | 5.1 | 8.2 | 6.9 | 12.1 | 6.9 | 11.3 |
| $Q = P$ | 4.9 | 7.6 | – | – | – | – | – | – | – | – | – | – |
| $Q = B$ | 4.8 | 7.6 | – | – | – | – | – | – | – | – | – | – |

Table 1: *The efficiency of expansion techniques is shown as the average query time in milliseconds over 100,000 queries on a machine with a large amount of memory (Lrg) and one with little (Ltl).* None *specifies the baseline,* QE *shows the standard local analysis results, and* $S$=40 *shows the results for a summarisation technique.* $Q$=$D$ *is the document centric expansion technique.* $Q \in V$, $Q$=$P$, *and* $Q$=$B$ *are term centric and phrase centric approaches.*

and *Ltl* respectively. *Lrg* has ample amount of memory that easily fits – at least for the experiments with small collections – the whole document collection as well as inverted indexes and any major auxiliary data structures, such as document summaries, if applicable. Even though main memory was flushed before timings were commenced, eliminating any influence of caching from any previous timed runs, as all 100,000 queries are processed, all data is eventually cached. The effect of this to the conventional QE method is that the additional time requirement over the baseline is purely that of parsing documents, evaluating terms and processing a greater number of inverted lists, rather than the main cost associated with expanding queries from a local set in a typical environment, which is retrieving documents from disk. In practice, for larger collections, this scenario is unrealistic. We therefore also show timings for a machine that can fit only part of the collection and inverted lists in main memory.

**Significance testing.** One cannot assume that two result sets differ significantly from each other by simply observing the magnitude of the difference of an evaluation measure over a number of queries. To evaluate our results we make use of the non-parametric Wilcoxon matched-pairs signed ranks test since it places no assumption on the distribution of test data. In particular, a non-parametric test does not require that data is normally distributed, which is important for our purposes.

## 5 Results

Results are listed in Tables 1 and 2. Only methods that were successful on our training data are reported. The $S = 40$ rows in Tables 1 and 2 give results for one of the most successful methods we previously explored when using in-memory document summaries [2]. For this method each document is summarised by the top 40 *tf.idf* terms of that document. During query time, the summaries for all documents are kept in memory. The parameter of $S = 40$ was not tuned for the WSJ2 collection. The memory overheads for this method are as follows: WSJ2: 11.7 MB, AP: 26.1 MB, NW: 84.0 MB, FBIS: 20.8 MB, FT: 33.2 MB, and LA: 20.4 MB.

Since the effectiveness of phrases experiments is no better than the other DE runs and the resource requirements are comparatively large for phrase experiments, we did not experiment with phrases further.

**Effectiveness.** In the following discussion we treat any sub-collection as a full collection and neglect a change of 0.005 or less in the respective measurements. Across 13 collections, MAP was increased ten times through QE and decreased twice, whereas the DE technique $Q = D$ improved only five collections and degraded the results of one. These figures are six and two respectively for the term centric method. QE improved precision at 10 in nine instances and decreased it in three cases. Retrieval results for precision at 10 were increased three times and decreased in five instances, employing either DE technique. Using those terms for comparison, the summarisation technique performs the same as QE, with the exception of FT where it is a little worse the QE. Increases in effectiveness for DE methods are small compared to those of QE. Furthermore, improvements achieved by QE are mostly statistically significant whereas DE improvements are not.

**Efficiency.** The term centric approach slows down retrieval by 2% in most cases, whereas the document centric technique adds roughly 10%. These figures are the same on both machines as there is enough main memory on either machine for caching of inverted lists.

On $Lrg$, QE slows down retrieval by a factor of five to seven. Caching does not work well, since during query evaluation many lists have to be purged in order to make room for other lists and for documents that are retrieved from disk. This problem is exacerbated on $Ltl$ where the overhead increases from five to fifteen-fold.

The additional data needed for the technique involving summaries fits well into memory on $Lrg$, while leaving adequate room for inverted lists to be cached. This is why query times are increased only by around 50%. On $Ltl$, some of the in-memory summaries need to be swapped in and out of memory more often and the penalty is relatively high, leading to a decrease in query throughput to roughly half of that of the baseline.

**Robustness.** Figure 3 shows how many queries are degraded or improved in respect to the baseline and by how much. The baseline is constructed by running queries in their original form against the non-modified corpus. All lines more or less intersect the x-axis at the same point, which means that all methods examined in this paper exhibit roughly the same robustness for each collection.

| Coll. | Method | | MAP | P@10 | R-Pr. | | MAP | P@10 | R-Pr. | | MAP | P@10 | R-Pr. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| WSJ2 | None | TREC 3 | 0.251 | 0.363 | 0.275 | | | | | | | | |
| | QE | | 0.325‡ | 0.388 | 0.324‡ | | | | | | | | |
| | $S = 40$ | | 0.286‡ | 0.380 | 0.287† | | | | | | | | |
| | $Q = D$ | | 0.265‡ | 0.361 | 0.280 | | | | | | | | |
| | $Q \in V$ | | 0.264 | 0.378 | 0.283 | | | | | | | | |
| | $Q = P$ | | 0.259 | 0.371 | 0.276 | | | | | | | | |
| | $Q = B$ | | 0.260 | 0.380 | 0.268† | | | | | | | | |
| AP | None | TREC 3 | 0.243 | 0.430 | 0.262 | | | | | | | | |
| | QE | | 0.327‡ | 0.468‡ | 0.333‡ | | | | | | | | |
| | $S = 40$ | | 0.290‡ | 0.454‡ | 0.301‡ | | | | | | | | |
| | $Q = D$ | | 0.251 | 0.416 | 0.286‡ | | | | | | | | |
| | $Q \in V$ | | 0.248 | 0.420 | 0.276‡ | | | | | | | | |
| NW | None | | | | | TREC 7 | 0.195 | 0.458 | 0.251 | TREC 8 | 0.222 | 0.438 | 0.262 |
| | QE | | | | | | 0.232‡ | 0.452 | 0.285‡ | | 0.250‡ | 0.464 | 0.289‡ |
| | $S = 40$ | | | | | | 0.208 | 0.438 | 0.263 | | 0.234 | 0.434 | 0.269 |
| | $Q = D$ | | | | | | 0.199 | 0.476 | 0.259 | | 0.213 | 0.444 | 0.263 |
| | $Q \in V$ | | | | | | 0.195† | 0.444 | 0.243 | | 0.220 | 0.434 | 0.261 |
| FBIS | None | TREC 6 | 0.223 | 0.260 | 0.232 | TREC 7 | 0.208 | 0.318 | 0.218 | TREC 8 | 0.269 | 0.319 | 0.281 |
| | QE | | 0.237‡ | 0.266 | 0.226 | | 0.222 | 0.292† | 0.243‡ | | 0.270 | 0.305 | 0.256 |
| | $S = 40$ | | 0.231 | 0.274 | 0.226 | | 0.217 | 0.308 | 0.224 | | 0.268 | 0.309 | 0.274 |
| | $Q = D$ | | 0.220‡ | 0.257 | 0.235 | | 0.205 | 0.300‡ | 0.218 | | 0.264‡ | 0.312 | 0.279 |
| | $Q \in V$ | | 0.233 | 0.260 | 0.237 | | 0.228 | 0.318 | 0.239‡ | | 0.278 | 0.321 | 0.284 |
| FT | None | TREC 6 | 0.214 | 0.250 | 0.244 | TREC 7 | 0.224 | 0.271 | 0.241 | TREC 8 | 0.290 | 0.331 | 0.298 |
| | QE | | 0.209 | 0.261 | 0.220 | | 0.233 | 0.287 | 0.234 | | 0.298 | 0.361† | 0.282 |
| | $S = 40$ | | 0.217 | 0.276‡ | 0.221 | | 0.216 | 0.269 | 0.229 | | 0.261 | 0.341 | 0.249 |
| | $Q = D$ | | 0.211† | 0.243 | 0.235 | | 0.229 | 0.277 | 0.242 | | 0.299 | 0.316 | 0.312 |
| | $Q \in V$ | | 0.206 | 0.237 | 0.229 | | 0.212 | 0.287‡ | 0.221 | | 0.295 | 0.325 | 0.304 |
| LA | None | TREC 6 | 0.198 | 0.231 | 0.232 | TREC 7 | 0.211 | 0.300 | 0.234 | TREC 8 | 0.233 | 0.260 | 0.238 |
| | QE | | 0.226‡ | 0.254‡ | 0.218 | | 0.251‡ | 0.316 | 0.269† | | 0.207 | 0.256 | 0.223 |
| | $S = 40$ | | 0.213‡ | 0.244‡ | 0.222 | | 0.240† | 0.306 | 0.263† | | 0.216 | 0.262 | 0.237 |
| | $Q = D$ | | 0.209 | 0.227 | 0.221 | | 0.225 | 0.304 | 0.242 | | 0.237 | 0.256 | 0.248 |
| | $Q \in V$ | | 0.216 | 0.237† | 0.237 | | 0.224 | 0.288 | 0.250 | | 0.235 | 0.256 | 0.242 |

Table 2: *Effectiveness of expansion techniques, averaged over 50 queries. The WSJ2 data was used for tuning. Shown are mean average precision (MAP), precision at 10 (P@10), and precision at the number of relevant documents (R-Pr.). Notation otherwise is the same as that used in Table 1. Results that are statistically significant different to the baseline at the 0.10 and 0.05 levels are marked with † and ‡ respectively.*

## 6 Analysis

An explanation for the relatively poor improvements of DE is that the topic of the expanded documents is changed too much from the original topic, analogous to query drift. This problem could be alleviated by adding a reduced weight to terms as they are added to documents. We leave this for future work.

A further explanation is that the lack of context during the expansion process is unhelpful; whereas, during conventional QE, several query terms set a particular context that determines the intersection of documents in the local set. Our experiments involving phrases try to address this problem. However, the generation method of phrases is most likely insufficient. Phrases are extracted from the collection itself – rather than from a suitable query log for example – and therefore no new context from outside the collection is found.

## 7 Conclusions

A series of experiments cannot prove that a family of methods is not viable. Establishing a positive result is straightforward; establishing a negative result involves demonstrating that all reasonable avenues of progress have been investigated and found wanting. Nonetheless, we believe we have shown that corpus-based DE is not promising. Other DE methods, based on extracting terms from external resources, have been found to give limited gains in some circumstances. However, while query-time costs are low, we were unable to use corpus-based DE to significantly improve effectiveness, and the index-time costs are considerable.

In contrast, our fresh investigation of QE showed that it was generally of benefit in the newswire collections used in our experiments, and that the evaluation costs can be much reduced while broadly maintaining the effectiveness gains. These results, we believe, should help focus future research in the area, by demonstrating that work on DE may not be warranted and by suggesting promising further directions for improving the efficiency and effectiveness of QE.

## Acknowledgements

We thank Nick Lester and William Webber. This research was conducted with the support of an APA, the State Government of Victoria, and an RMIT VRII grant.

## References

[1] B. Billerbeck, F. Scholer, H. E. Williams and J. Zobel. Query expansion using associated queries. In *Proc. Int. Conf. on Information and Knowledge Management*, pages 2–9, New Orleans, LA, November 2003. ACM Press, New York.

[2] B. Billerbeck and J. Zobel. Techniques for efficient query expansion. In A. Apostolico and M. Melucci (editors), *Proc. String Processing and Information Retrieval Symp.*, pages 30–42, Padova, Italy, September 2004. Springer-Verlag.

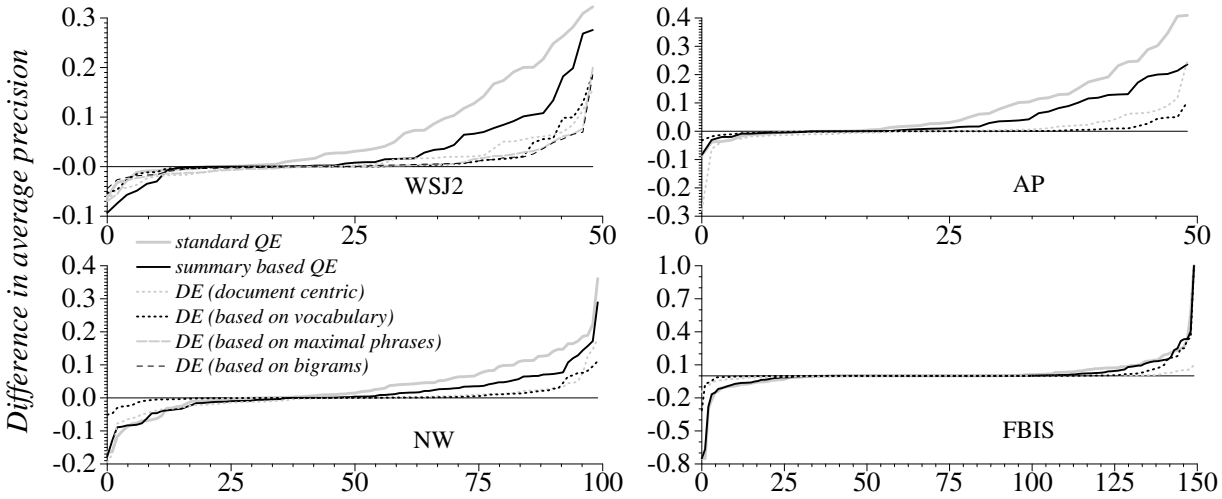



Figure 3: *The graphs show per-query-differences in average precision between each of the methods and the respective baselines. Each curve for each collection is sorted individually. Data for phrases and bigrams is only shown for WSJ2. The graphs for FT and LA are similar to that of FBIS and are not shown here.*


[3] W. B. Croft (editor). *Advances in Information Retrieval*. Kluwer Academic Publishers, Norwell, MA, 2000.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas and T. Landauer. Indexing by latent semantic analysis. *Jour. of the American Society for Information Science*, Volume 41, Number 6, pages 391–407, 1990.

[5] E. Ide and G. Salton. Interactive search strategies and dynamic file organization in information retrieval. In G. Salton (editor), *The SMART Retrieval System – Experiments in Automatic Document Processing*, pages 373–393. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[6] N. Lester and H. E. Williams. Topic tracking at RMIT University. In *Topic Detection and Tracking Workshop (TDT)*, Gaithersburg, MD, 2002. National Institute of Standards and Technology.

[7] G.-A. Levow and D. W. Oard. Signal boosting for translingual topic tracking: Document expansion and n-best translation. In *Topic Detection and Tracking: Event-Based Information Organization*, pages 175–195. Kluwer Academic Publishers, 2002.

[8] Y.-C. Li and H. M. Meng. Document expansion using a side collection for monolingual and cross-language spoken document retrieval. In *ISCA Workshop on Multilingual Spoken Document Retrieval*, pages 85–90, Hong Kong, 2003.

[9] L. A. F. Park and K. Ramamohanarao. Hybrid pre-query term expansion using latent semantic analysis. In *Proceedings of the fourth IEEE International Conference on Data Mining (ICDM'04)*, pages 178–185, Washington, DC, USA, 2004. IEEE Computer Society.

[10] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Jour. of the American Society for Information Science*, Volume 27, Number 3, pages 129–146, 1976.

[11] S. E. Robertson and S. Walker. Okapi/Keenbow at TREC-8. In E. M. Voorhees and D. K. Harman (editors), *Proc. Text Retrieval Conf. (TREC)*, pages 151–161, Gaithersburg, MD, November 1999. National Institute of Standards and Technology Special Publication 500-246.

[12] J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton (editor), *The SMART Retrieval System – Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[13] I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In J. Callan, G. Cormack, C. Clarke, D. Hawking and A. Smeaton (editors), *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*, pages 213–220, Toronto, Canada, July 2003. ACM Press, New York.

[14] F. Scholer, H. E. Williams and A. Turpin. Query association surrogates for web search. *Jour. of the American Society for Information Science and Technology*, Volume 55, Number 7, pages 637–650, 2004.

[15] A. Singhal and F. Pereira. Document expansion for speech retrieval. In F. Gey, M. Hearst and R. Tong (editors), *Proc. ACM-SIGIR Int. Conf. on Research and Development in Information Retrieval*, pages 34–41, Berkeley, CA, August 1999. ACM Press, New York.

[16] K. Sparck Jones, S. Walker and S. E. Robertson. A probabilistic model of information retrieval: Development and comparative experiments. Parts 1&2. *Information Processing & Management*, Volume 36, Number 6, pages 779–840, 2000.

[17] A. Spink, D. Wolfram, Major B. J. Jansen and T. Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, Volume 35, Number 3, pages 107–109, March 2002.

[18] K. Taghva, J. Borsack, T. Nartker and A. Condit. The role of manually-assigned keywords in query expansion. *Information Processing & Management*, Volume 40, Number 3, pages 441–458, 2004.

[19] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, Volume 18, Number 1, pages 79–112, 2000.

# ePOC: Mobile Clinical Information Access and Diffusion in Ambulatory Care Service Settings

*Peter Eklund*
School of Economics and Information Systems
University of Wollongong
NSW, 2522, AUSTRALIA
peklund@uow.edu.au

*Jason Sargent*
School of Economics and Information Systems
University of Wollongong
NSW, 2522, AUSTRALIA
jsargent@uow.edu.au

**Abstract** *This paper represents a preliminary overview (work-in-progress) of a mobile e-Health research and development project and the intrinsic considerations which arise when designing such patient data management systems tailored to ambulatory care. Its purpose is to give an outline of the issues that allow technological enablement of electronic patient data management in the delivery of home-based medical care. While the replacement of more traditional paper-based patient data management using Personal Digital Assistants as a collection platform is technically straightforward, the organizational realignment of an electronic document management system requires careful study and deployment in order to maximize success. We outline the methodological considerations for document management diffusion within this e-Health setting and describe the issues, architecture and proposed rollout of an electronic Point-Of-Care (ePOC) system.*

**Keywords** e-Health, document management and workflow, information access and diffusion

## 1. Introduction

Hospital in the Home (HITH) patients are those who without the provision of the hospital in the home service would require inpatient care by the nature of their medical or social condition [5]. Traditional community-based healthcare services such as HITH and Ambulatory Care are based on a healthcare delivery model of providing episodes of care to decentralised hospital outpatients; in the patient's own home or aged-care facility. However, the information management model utilised by HITH and Ambulatory Care service providers such as The Ambulatory Care Team (TACT) Northern Illawarra is centralised. A patient's electronic medical record (EMR) is only accessible by a TACT clinician while the clinician is located within the wired hospital network architecture, yet the treatment of the patient (based on reference to the EMR) occurs away from such wired architecture. This misalignment of models is at best problematic by creating inefficiencies and duplication in regards to clinical information access and diffusion.

A hybrid information system of paper and electronic formats exists for Ambulatory Care. After referral of a patient, the system process begins with downloading the patient's EMR (a task performed by TACT administration personnel). The EMR is printed and appended to an assortment of paper-based forms that are utilised during the patient visit for data collection pertaining to the patient's episode of care (treatment). Finally, upon return to the office, post-visit clinician documentation is forwarded to a data entry clerk for transcription and the process of electronic upload of the updated patient record to an appropriate centralised patient/ hospital database occurs. A complete paper record however, is also stored on-site at the TACT office. This document workflow therefore corresponds to a paper-based model of information management that lends itself to a more integrated electronic document management system.

In a hybrid system of this sort, alignment is needed between the models of *service* and *information management* for mobile community-based healthcare to occur "...[for] ICT to improve the delivery of healthcare and drive efficiencies through the health sector" [3]. As early as 2001, the New South Wales Department of Health recognised effective healthcare delivery within community-based health services depends on efficient information access. [15]. More recently, the Australian Government flagged e-Health as a

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005. Copyright for this article remains with the authors.**

priority area, noting that ICT will be integral to the strategy of reforming and improving delivery of healthcare in Australia [3]. Leveraging the utility of mobile devices such as Personal Digital Assistants (PDAs) in addressing these needs requires several alignments for a project such as this to succeed: (i) alignment between the legacy document management and new technology, (ii) organisational realignment of the technology to institutional and legal collection requirements (iii) sensitivity of the electronic document collection and its operators, in this case Health care workers.

## 2. Legacy Document Management

A systematic approach must be taken to transforming paper-based systems to electronic systems. In the context of Ambulatory Care, this systematic approach includes setting a baseline for capturing clinical data requirements by adhering to minimum data sets. Further, noting points of data collection and diffusion along clinical pathways helps to identify feeder systems (many of which are legacy systems).

### Minimum Data Sets

An important consideration in designing mobile-based information systems for Ambulatory Care is the capture of appropriate, and necessary, data. Information systems (both electronic and paper-based) for Ambulatory Care are built around minimum data sets. Prior to the development of the Victorian HITH Minimum Data Set, data sets focused upon either inpatient areas or community-based care. Therefore, the individual data sets did not meet the requirements of HITH programs, as these programs cross care boundaries from the acute inpatient setting to the community [1].

### Feeder System Integration

Integration of any proposed electronic system will centre on linking of records to clinical guidelines and protocols if “best-practice is to be embedded as an integral part of the health care delivery process” [6]. Alignment between legacy document management and new technologies is best achieved through an investigation of feeder systems on which any new technology device (PDA) will rely upon to populate fields in a patient’s consolidated health record. Investigations have been carried out to determine what databases and which NSW Health information systems contain the data required to provide the PDA with necessary patient and clinical record information. In addition, links between NSW and regional health information systems have been studied. The outcome of this analysis shows that the Community Health Information Management Enterprise (CHIME) system holds the registration details for TACT clients and is the main information patient demographic system that TACT interacts with.

Other related patient data is also accessed from other information sources, including medications, pathology and radiology results held in other health information systems within the regional and state-based health system. However, both the present and future CHIME-based data collection is intended to remain paper-based for the ambulatory care health workers in the foreseeable future. Therefore, a PDA-based collection system has natural advantages to the existing health delivery and information acquisition regimes.

The PDA-based clinical recording system for ambulatory health workers confronts its main difficulty in terms of its integration with existing NSW Health and regional information systems and their data. Therefore the problems for developing electronic Point-Of-Care systems are twofold (i) to obtain the authority to access appropriate information and (ii) to electronically return the information collected from the mobile system to the feeder health information systems. In both instances, the person(s) performing data collection and dissemination roles in the TACT process will not change when TACT paper-based systems transition to an electronic Point-Of-Care (ePOC) system. The only differentiation will be in the data format; digitized compared to hand written or typed. This is anticipated to alleviate any ‘authority to access’ issues.

### Messaging

Connection of distributed e-Health systems requires support by generic middleware components, while interoperability is addressed by messaging. A Health Level 7 (HL7) messaging gateway handles messaging from a clinical trial server to the PDA. HL7 is an ANSI-accredited standards developing organization “dedicated to providing a comprehensive framework and related standards for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services” [8]. HL7 Version 3 will be a key part of the contribution of IT to healthcare’s reaching new levels of (1) effective and cost-efficient patient care decisions, (2) safety and cost savings that come from ‘doing it right,’ in the sense of preventing avoidable errors, and (3) the aggregation of health information for evidence-based medicine and databased policy [19].

Once appropriate applications are implemented for TACT, (such as CHIME appointments and similar) messaging is intended to be switched to use these systems and the initial clinical server will be retired. ePOC messaging will conform to HL7 Version 3, which initially will use only XML encoding [10]. The XML tags in a Clinical Document Architecture (CDA) document are defined by the HL7 Reference Information Model (RIM) which is based on a variant of Unified Modeling Language (UML).

CDA is based on the RIM and uses V3 data types and methodology. It contains many optional data elements and data segments, making it adaptable to almost any site. This feature makes the decision to base messaging on HL7 V3 attractive to the ePOC project as it complements the ethos of the project in developing an application which is generic, scalable and adaptable. Further, "the Reference Information Model (RIM) is the cornerstone of the HL7 Version 3 development process. RIM is a large pictorial representation of the clinical data (domains) explicitly representing the semantic and lexical connections that exist between the information carried in the fields of HL7 messages" [9].

Apart from technical data management issues associated with transforming paper-based point-of care systems to electronic point-of-care systems, social issues (ethics, privacy and security) also exist which must be addressed in regards to the realignment of any newly proposed technology implementation at the organizational level.

## 3. Implicit Information Processes - Organizational Considerations

The mode of data access, collection and diffusion by TACT clinicians at point-of-care is obviously transformed by the integration of mobile devices such as PDAs. For example, recording a patient's blood sugar level (BSL) by entering the BSL directly into a preformatted field on a PDA differs from the present mode whereby the information is written into a field on a paper form. What is not so obvious, however, is the implicit change in the process of data collection and diffusion from an ethical and patient privacy perspective.

Permission for an Ambulatory Care clinician to access, modify and update patient information as part of delivering an episode of care currently (existing paper-based system) is deemed to have occurred when a patient agrees to be treated at point-of-care (i.e. the patient's residence or Ambulatory Care outpatient facility). This occurs as a subsystem of the referral process. The ethical connotation as relates to patient permission for a TACT clinician to use a non standard device (PDA) to access and modify their (patient) records was not predicted by ePOC systems designers at the time.

Upon reflection and advice sought under guidance from an overseeing Human Research Ethics Committee, ePOC developers became aware of such differentiation. A trail must be blazed by ePOC systems and it is the ePOC project team's belief that patient acceptance of PDA-based patient document management systems will gain acceptance by patients as they become accustomed to the technology and the benefits of deploying such devices, leading to improved levels of care, are made known through exposure to the new paradigm of ePOC systems.

This sampling of issues associated with the need to realign new technologies and processes with existing legacy document management and embedded institutional ethical/ legal information collection illustrates the complexity of developing mobile health information systems for Ambulatory Care. A field trial which attempts to proactively address these issues in a pragmatic manner is the ePOC PDA Project, currently under development for TACT (ePOC Client).

## 4. *e*POC: A Mobile e-Health Solution

Paper-based information available to a clinician at point-of-care is effectively limited to what the clinician is able to carry. The ability for the TACT clinician to electronically modify a patient record 'in the field' is not possible at present. Additionally, problems associated with paper-based exchanges, difficulty in deciphering hand-writing, lack of integration of information and limited availability and capture of information at the point-of-care have been identified [16]. *Electronic* Point-of-Care offers the potential to overcome these identified limitations of paper-based Ambulatory Care systems.

The ePOC PDA project is a collaborative research and development project between academic, health and health informatics partners. Researchers are drawn from three Australian universities; Wollongong University, Flinders University and the University of South Australia. South Eastern Sydney and Illawarra Area Health Service (SESIAH) perform the role of health partner, while Pen Computer Systems Pty Ltd, a leading Australian-based health informatics company, performs the role of technical partner.

The convergence of information and communication technologies (ICTs) into a single

mobile device (PDA) may well be seen as a harbinger of viable mobile e-Health solutions for community-based healthcare services. PDAs as platforms for mobile-based *hospital* clinical information have "proven to be among the most cost effective ways to improve patient care quality and reduce medical data collection errors" [13]. Studies of such usage within hospital environments are plentiful [2, 11, 12, and 13]. However, studies which explore extending PDA usage into Ambulatory Care service settings, as an *electronic* point-of-care application are scant.

With the process of digitization, any information can be delivered through any medium with the user deciding what form it takes [17]. A PDA-based point-of-care system is significant because it provides for the collection, delivery and exchange of timely information (both text and images) at the point-of-care. Such a system provides natural advantages over the existing paper-based clinical and administrative collection systems. The Health Informatics Society of Australia broadly defines health informatics as "an evolving socio-technical and scientific discipline that deals with the collection, storage, retrieval, communication and optimal use of health related data, information and knowledge" [7]. The ePOC PDA project is in effect, an applied paradigm of health informatics. The broader aim of the ePOC system is to become the archetype, a prototype for the access and diffusion of clinical Ambulatory Care data at point-of-care.

TACT has a requirement to enter data at the point-of-care about a patient condition and the clinical activities carried out during the ambulatory visit. This will be provided by an application on a PDA that synchronises with the required health information database(s). The evaluation of the PDA system in the ambulatory care environment provides a central platform for the research and the measurement of its impact on ambulatory clinical practices.

The ePOC information solution as illustrated in Figure 1 consists of four steps:

1. Query to PIMS (Patient Information Management System / HOSPAS Equivalent) for all Patients discharged in the last 24 hours
2. Demographic data for Patients returned via e*Gate
3. Appropriate Patient information is sent via the HL7 Messaging Gateway to the PDA
4. Updated Patient information is sent back to ePOC via the HL7 Messaging Gateway

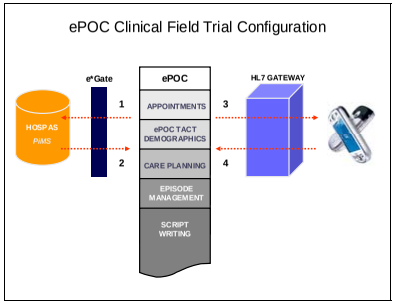


Figure 1: Illustrated ePOC Information Solution

## 5. Project Methodology

ePOC is a multi-phase, iterative R&D project with a research focus including, but not limited to: Pre and Post-Implementation technology implementation framework analysis, a detailed clinical workflow study, the investigation of end-user perceptions of clinical mobile health information systems, the iterative evaluation of user acceptance by the client (TACT), and the exploration of technical issues involving human computer interface development and evaluation, database schema browsing, document management, rapid application development and human-centred design.

An initial ePOC feasibility study of a prototype PDA based health information system for TACT was conducted in 2003 [18]. The identified benefits of this study included: supporting critical workflow activities, addressing operational inefficiencies (by automating and streamlining workflow activities) and increasing access to patient and medical information at the point-of-care. Additional requirements were also noted including: data entry audit controls, patient allergy reminders, data formatting standardisation, patient record search (database query), GUI design, data sharing interface, patient notes, controls and search mechanisms, usability issues, and changing and unpredictable datasets [18]. In its implementation, ePOC addresses the difficulties of federating disparate data, HL7 messaging and the limitations of the mobile platform within a practical community clinical health service environment. ePOC is intended to address the mobile clinical information needs for TACT clinicians and allied-health professionals.

**Electronic Document Collection: End-User Implications**

In order to achieve best fit between a PDA-based health information system and end-user (clinician), issues of user acceptance must be identified and mitigated with change management strategies. An assortment of existing paper forms utilized for patient data collection can easily be mirrored as proformas for deploying on a PDA in electronic format. An associated process however, was a review of these current forms (their structure, data flow, legal data gathering requirements and data field integration into appropriate back-end health systems. In this way, the electronic versions (proformas) are deemed to be the most appropriate data collection method at the time. This approach can then be refined during the course of clinical trials before the delivered system becomes operational. An electronic information system replica for TACT could be developed in a short time. Such a situation should minimise the risk that the desired delivered system will not meet organizational, operational or more importantly, end-user requirements.

Anecdotal evidence abounds of poorly planned systems integration projects which are doomed to failure. The ethos of the ePOC PDA project is that with forward thought and proactive approaches (including iterative consultation) instigated by the systems development team, user acceptance can be improved and ownership of the delivered system taken by clinicians as a result of being involved in a nontrivial manner during all phases of system development. Additionally, this approach helps address the issue of the varied computing skill levels of the TACT members.

A combination of structured methodological approaches including focus groups, information and question and answer sessions were conducted during the pre-implementation phase of ePOC. End-users gained an understanding of the proposed system by interacting with the ePOC prototype, deployed during an initial feasibility phase of the project. Concerns regarding not having a complete working system to base perceptions upon are not an issue for the ePOC system developers. Developers instead followed Davis and Venkatesh's hypothesis "that stable and representative measures of perceived usefulness only require that potential users be informed of what a system will be designed to do, i.e., its intended functionality, and do not require hands-on interaction with a working system" [4].

It is evident ePOC will transform the workflow for TACT clinicians at point-of-care, however, administration and management personal also stand to benefit from ePOC by the utility of data the system will provide in relation to organizational and operational statistics and integrated reporting. These components of TACT are performed manually, in a time consuming manner, at the end of each month or specified reporting period. On occasion, reports are requested by Area Health Service managers on an ad hoc basis. ePOC can satisfy such requests in a far timelier manner than would presently be possible by administration personnel.

## 6. Conclusion

This paper has outlined several intrinsic considerations which arise in the development and proposed subsequent integration of PDA-based mobile document management systems designed for deployment within Ambulatory Care service settings. Many of the issues described manifest themselves as a result of the misalignment in *service* and *information management* models utilised by community health services such as Ambulatory Care and HITH.

Health systems by their very nature are fundamentally dynamic in nature, evolving over time, consisting of a mixture of legacy, integrated and stand alone systems; the development of which are constrained by the need to extend resources under tightening budget constraints as the Australian population ages and further burden is placed upon existing health services. A PDA-based e-Health solution which achieves a 'best fit' for TACT clinicians by addressing the alignment of *service* and *information management* models of community health care leverages ICT to improve clinical information access and diffusion under such identified constraints.

## 7. References


[1] Ambulatory Care Australia. *The Victorian HITH Minimum Data Set*. [Online] Available <URL: http://www.health.vic.gov.au/aca/mds.htm > Last accessed 21/09/2005.

[2] Ammenworth, W. et al. Mobile Information and Communication Tools in the Hospital. *International Journal of Medical Informatics*. 57, pages 21-40, 2000.

[3] Coonan, H. *New priorities in ICT*. [Online] Available <URL: http://www.minister.dcita.gov.au/media/speeches/new_priorities_in_ict__opening_of_future_parc__cebit_australia_sydney__wednesday_25_may_2005 > Last accessed September 27, 2005.

[4] Davis, F, and Venkatesh, V. Toward Preprotoype User Acceptance Testing of New Information Systems: Implications for Software Project Management. *IEEE Transactions on Engineering Management, 51(1)*, February 2004.

[5] Grayson, L. *What is Hospital in the Home?* Australian Home and Outpatient Intravenous Therapy Association Annual Scientific Meeting - Conference Proceedings; Australian Home and Outpatient Intravenous Therapy Association, 1996.

[6] Grimson, J., Grimson, W., and Hasselbring, W. *The SI challenge in health care*. Communications of the ACM, Vol 43 Number 6, pages 48-55, 2000.

[7] Health Informatics Society of Australia. *What is Health Informatics?* [Online] Available <URL: http://www.hisa.org.au/100101.php > Last accessed 18/09/2005.

[8] HL7. *What is HL7?* [Online] Available <URL: http://www.HL7.org/ > Last accessed 28/09/2005.

[9] HL7 RIM Reference Information Model. [Online] Available <URL: http://www.hl7.org/about/hl7about.htm#RIM > Last accessed 6/10/2005.

[10] HL7 CDA Clinical Document Architecture. [Online] Available <URL: http://www.hl7.org/about/hl7about.htm#CDA > Last accessed 4/10/2005.

[11] McCord, L. Using a personal digital assistant to streamline the OR workload. *Journal of Association of periOperative Registered Nurses*. December 2003. [Online] Available <URL: http://www.findarticles.com/p/articles/mi_m0FSL/is_6_78/ai_111895687 > Last accessed 12/09/2005.

[12] McCreadie, S.R., Stevenson, J.G., Sweet, B.V., & Kramer, M. Using Personal Digital Assistants to Access Drug Information, *Am J Health-Syst Pharm*, 59(15), pages 1340-1343, 2002.

[13] MedPDA.Net. *Doctors using PDAs make less mistakes*. Medical PDA reviews. [Online] Available <URL: http://www.medpda.net/archives/001050.html > Last accessed 27/06/2005.

[14] Mitchell, J. (1999). *From Telehealth to E-health: the Unstoppable Rise of E-health*. [Online] Available <URL: http://www.dcita.gov.au/ie/publications/1999/09/rise?SQ_DESIGN_NAME=printer_friendly > Last accessed September 30, 2005.

[15] NSW Health. Information Management Division. *Information management and Technology, Strategic Plan*, December, 2001.

[16] Soar, J & Yu, P. *The Future of eHealth through next generation Wireless and Mobile Connectivity*. Proceedings Health Informatics Conference, HIC2002, Melbourne, August, 2002.

[17] Turner, C. *The information e-conomy: business strategies for competing in the global age*. London: Kogan Page, 2002.

[18] Walsh, D., Alcock, C., Burgess, L., & Cooper, J. *PDAs and Effective Community Healthcare Delivery: a Mobile Technology Solution to Point-of-Care Health Services Delivery for Ambulatory Care*. Proceedings from CollECTeR LatAm, 2004.

[19] XML Coverpages. *HL7 Announces ANSI Approval of Several Health Level Seven V3 Specifications.* [Online] Available <URL: http://xml.coverpages.org/ni2004-07-21-a.html > Last accessed 2/10/2005.

## 8. Acknowledgement

ePOC is under development for The Ambulatory Care Team (TACT), Northern Illawarra. The ePOC PDA Project is an Australian Research Council (ARC) Linkage Grant with collaborative funding, research and development by the University of Wollongong, South Eastern Sydney & Illawarra Area Health Service (SESIAHS) and Pen Computer Systems Pty Ltd.

# An Experimental Study of Workflow and Collaborative Document Authoring in Medical Research

*Venkata Nallaparaju*[†] *Gitesh K. Raikundalia*[†] *Caroline Brand*[§] *Christopher Bain*[§] *Ana Hutchinson*[§]

[†]School of Computer Science and Mathematics
Victoria University
PO Box 14428
Melbourne City MC 8001 Australia

[§]Clinical Epidemiology and Health Service Evaluation
Melbourne Health
Parkville, Melbourne 3050 Australia

*vgd28 @yahoo.com* *Gitesh.Raikundalia @vu.edu.au* *Caroline.Brand @mh.org.au* *Christopher.Bain @mh.org.au* *Anastasia.Hutchinson @mh.org.au*

**Abstract** *Workflow is asynchronous technology widely used in the automation of organisational processes. Workflow provides benefits such as greater efficiency in an organisation, better worker productivity and greater process control. Synchronous collaborative authoring tools are technologies that allow a group of dispersed authors to write a document at the same time. These tools are beneficial in assisting authors to write some proportion, if not all, of a document simultaneously.*

*This paper presents findings from an experiment combining both workflow and collaborative authoring tools in a medical research environment. Studies investigating the combination of these tools are few, resulting in a lack of understanding of how this combination can effectively assist organisations in document-based processes. Overall, the combined workflow/collaborative authoring solution was found effective in the generation of a medical research paper.*

**Keywords** Workflow, collaborative document authoring, medical research, experimental study.

## 1 Introduction

The Workflow Management Coalition (WfMC) is a major international organisation that promotes workflow by development of standards for workflow systems. The WfMC defines a workflow system as a system "that defines, creates and manages the execution of workflows through the use of software, running on one or more workflow engines" [1] where a workflow is the automation of a process. Numerous workflow systems have been developed over several years for the purpose of assisting humans in business and other processes. For instance, the processing of a home loan application in a bank is an obvious multiple-stage problem where automation of the process with technology assists greatly. The workflow allows the bank to track the required application documents and forward them to the appropriate staff members, thereby making the whole process more efficient.

Collaborative authoring tools are tools that allow multiple dispersed users to create and work on a document simultaneously. The benefit of such a tool is that members of a group may collaborate on the document from their own computers. Also, a group member does not have to wait for another member to finish their current work on the document before they can contribute to the document. The members can work efficiently on the document at the same time, assisting each other in writing its contents. For instance, two university lecturers could write and finish an assignment handout (specifications of the assignment) together simultaneously, from their own computers.

A combination of workflow and collaborative authoring tools is worth investigating partly because of the benefits provided by the tools separately from one another. However, these tools are also worth investigating in their operation together. This is because together they form the exact software solution to automating a document-based process

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

where multiple authors may wish to work jointly and simultaneously.

Documenting and publishing of findings in medicine is a process that is often highly collaborative, involving many medical researchers, often taking many stages to complete. Given there are many stages in the process, where at least some of the stages would require more than one researcher to cooperate at the same time, this process can benefit highly from automation. Automation of this process should reduce human intervention as much as possible. For instance, researchers do not have to email the research paper to one another, and do not have to keep track of who has to work on the paper at which stage. Clearly, workflow tools are designed to launch either single-user or multiple-user tools as required by users. This project explores the effectiveness of collaborative word processing in a document-based process automated by workflow.

The combination of workflow and collaborative authoring tools in the support of document authoring has not been addressed sufficiently. A small number of systems exist that support collaborative creation of content assisted by workflow. For instance, Ho, Leong and Lam [2] describe a CORBA-based workflow framework that integrates a collaborative editor. Therefore, to the authors' knowledge, there is indeed a tremendous lack of research regarding experimental results for this combination in a real-world setting.

This paper describes an experiment on the combination of workflow and collaborative authoring in the generation of a medical research paper by a team of medical researchers. The medical researchers are from *Anonymous Medical Organisation* (AMO) in *City*, *Country*. The researchers work in an environment in which research papers and reports are frequently under development and require iterative input from multiple colleagues. Traditionally, group members work on the document contents separately. They may work on different sections of a document or on the same sections of the document, but will carry almost all of this out at different times from one another. Thus, one researcher is required to merge all contributions, once they have all been received, at the end of the entire process. Since documents often undergo a number of drafts, this effort can be time consuming and laborious as well as confusing.

The medical researchers have not been exposed to automating this document production process using workflow. Nor have they been exposed to the use of a collaborative word processor in enabling them to work together on the document simultaneously. This experiment required the researchers to write a research paper using the collaborative word processor, *CoWord* [3], and the workflow tool, *TrackNShare* [4], was used to automate the entire paper generation process.

For space reasons, screen captures of TrackNShare and CoWord have not been included in this paper. However, the CoWord user interface is very easy to understand: it is the same as the user interface of Microsoft Word since CoWord operates by using the version of Microsoft Word on the user's own computer! The purpose of the CoWord system is to take over the user's own installation of Microsoft Word and make it collaborative. However, for certain technical reasons known to the developers of CoWord, the software does not provide all Microsoft Word functionality when Microsoft Word is made collaborative (e.g., full table creation functionality is not available). TrackNShare operates generally in the same way as standard centralised workflow tools.

Timely and well-written medical research papers contribute to dissemination and uptake of new information that is of considerable public benefit. A research paper is highly relevant to this research based on workflow and collaborative authoring because:

- A research paper is a frequently developed document of considerable importance to medical researchers who wish to disseminate their ideas and research findings and improve their career prospects. The researchers in this experiment have experience in producing this document type, and therefore, the researchers' comments in this experiment are appropriate and substantiate the findings of this paper.
- A research paper requires brainstorming of ideas, as attested by the medical subjects, which is one of the task types that collaborative editors and word processors are designed to support.
- Research paper authoring is collaborative and is a process that can be automated naturally with workflow.

The medical researchers who participated in this experiment need to publish their work in various journals, such as Health Services Research or the Journal of Health Services Research & Policy, and conferences, such as the Health Services & Policy Research Conference or the International Conference on the Scientific Basis of Health Services. The research paper the subjects wrote in this experiment was a conference paper about approaches and techniques to ensure that patients interact successfully with a health care system. The researchers were particularly interested in writing about successful interactions with hospitals, and are aiming to submit the paper to one of the above conferences.

The number of medical researchers (experimental subjects) in the AMO team is three. Hence, the group size of three in this experiment is determined by the number of members of the team, which is clearly also the number of co-authors that worked on the experimental paper together. Consequently, this research project investigates support of a small group,

and small groups are the group size usually expected to use a collaborative word processor or editor.

## 2 Related work

This section covers related work regarding the types of technologies that are used in this project. The three forms of related work cover workflow, collaborative authoring, and lastly, combined workflow and collaborative authoring.

Various types of workflow systems exist and are widespread in use. Web-based workflow tools have increased in number since the rise in popularity of the Web. IBM Lotus Domino [5] is a well-known, commercial system incorporating different collaboration technologies, including workflow. The most recent versions of this system provide Web user interfaces for using the system. WWWorkflow [6] is a system that is distinguished by its "careful separation of process mediation from product data management". Fakas and Karakostas [7] present peer-to-peer technology for managing dynamic workflow using Web Workflow Peers.

Another type of workflow is that of component-based systems. A generic workflow framework, BPAFrame, is presented in [8] that uses business objects in modelling processes and resources. WASA2 [9] supports flexible workflows in a heterogeneous environment and is built from the CORBA framework. Yongyi and Weishi [10] describe the component-based architecture of BetterProcess, which is a distributed software process management system.

Apart from the Web and components, other bases and perspectives exist from which workflow systems are developed. However, there are too many bases and perspectives to cover here. Nevertheless, workflow systems and technologies have indeed been used widely in various different domains. For instance, domains such as banking [11], law [12] and pharmacy [13].

A number of collaborative authoring systems have also been produced over the last few decades (although there are not as many of these systems compared to workflow systems). Systems include those such as SASSE [14], JAMM [15] and MoonEdit [16]. In order to implement such systems, certain issues have had to be addressed. An example of an issue is that of inconsistency. Inconsistency can occur in the form of divergent results where a document is replicated at different sites. A solution to this inconsistency problem is a consistency model as explained in [17].

Some systems may have specific applications rather than being general document editors. Clay [18] is a collaborative environment that allows geographically distributed software developers to work together synchronously. Qingzhang, Zangyin and Kezhen [19] present work on simultaneous collaboration on XML documents.

As stated earlier, very few systems exist that automate a document-based process using workflow whilst supporting collaborative authoring. This may likely have been because any relevant tool can be integrated with a workflow system. Thus, apart from the related work covered next, there may have been no great need to specifically integrate a collaborative authoring tool than any other tool. Workflow systems are usually generic and allow use of any appropriate tool during task execution. However, experimental results for combined workflow and collaborative authoring are lacking. The following will therefore cover the few known systems that combine both technologies. Workflow systems that support

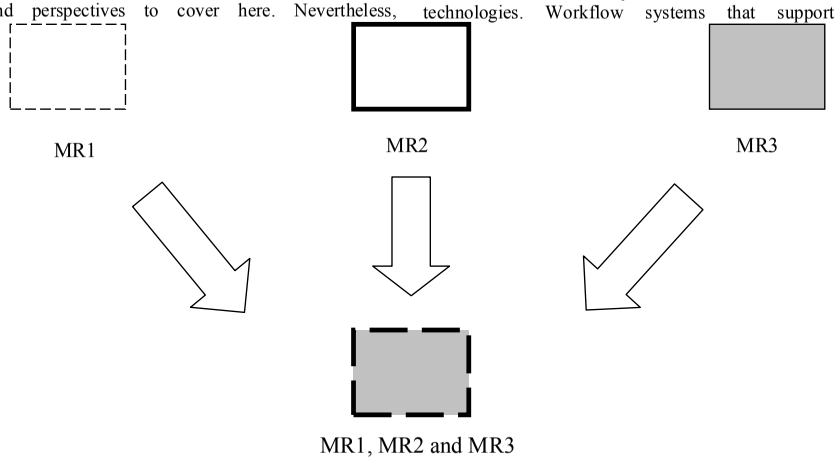



Figure 1: Merging of documents into final paper

asynchronous collaboration only are not covered here.

Ho, Leong and Lam [2] present a system where the documents a group works on are converted into XML format. Storage of general document content in XML format means that the attributes of XML can be applied to assist in document processes. For instance, the fragment nature of XML assists in access control and resource locking of contents. [20] describes a prototype, WoTel, that integrates a multimedia collaboration system to allow dispersed users to conduct audio/video conferencing whilst working on documents. Hodel, Gall and Dittrich [21] describe their TeNDaX architecture that supports synchronous editing, but stores the document contents within a database unlike approaches of other research. The researchers cover an evaluation of their system. However, this is not a trial of any form from which results can be applied to combined workflow and collaborative editing. Finally, Joeris [22] explains the use of synchronous collaboration and workflow in support of engineering domains.

## 3 Experimental Design

Figure 1 shows the current procedure in AMO for collaborative authoring of research papers. This is the way in which AMO researchers, *MR1*, *MR2* and *MR3*, have always authored a paper before their involvement in the experiment. The word processor used throughout is Microsoft Word. The Figure shows that medical researchers work on their own contributions to the paper independently of one another. For instance, MR1 will work on the abstract, introduction and methodology sections of the paper. Once all researchers have completed their contributions to the paper, the separate contributions are merged together into the final version. The researchers will work together *on the same computer to finalise the paper*. In this case, all three researchers will be "huddled around" one researcher's computer, discussing the paper.

Figure 2 shows the experimental workflow, configured with TrackNShare, which the researchers followed in authoring a research paper using CoWord. This was the workflow-driven form of the collaborative authoring process we were exposing the researchers to for the first time in their experience. Subjects sat at their own work computers at the AMO site. Unlike in Figure 1, there is simultaneous collaboration on the document, using CoWord, almost all the way throughout the entire process. The first task is where MR1 initiates the paper (using either CoWord as a single user or Microsoft Word). The AMO researchers felt that in this new configuration it was more effective for one of them to begin the paper so that others can work on the paper later—there was simply no need for more than one researcher to begin the paper together. Apart from the first task, all tasks in the workflow involved two or three researchers working on the paper at the same time using CoWord. Subjects spent half an hour on each of the tasks as shown in Figure 2.

Before the experiment, the subjects decided on the paper they would write in the experiment, through simple, informal, verbal discussion. The subjects discussed the content of the paper, deciding who would contribute which content to the paper. They also decided who would be involved in which of the tasks shown in Figure 2. The conference paper the subjects wrote was, "Engineering a Safe Landing: attitudes, knowledge and participation of medical clinicians in organisational patient safety systems". This paper was about patients' potentially successful interactions with health care systems, particularly hospitals. When the experiment finished, the paper ended up being just over six pages in length.

A subject would use a specific TrackNShare user interface to forward the document to the next subjects. When subjects were configured to work next on the document in the workflow, TrackNShare informed them by presenting a special user interface. One of the subjects would open the document, and all that the other subject(s) had to do was to use a specific CoWord window to join the session. The other subject(s) collaborating simultaneously with the first subject would therefore view the same document on their screen(s) in CoWord. The subjects would work on the document and one of them would forward the document onto the next subjects. However, the subjects were allowed no forms of communication at all (such as telephone, chat/instant messaging tools, etc.) with one another. It was hoped that the subjects may flag any difficulties they found in collaboration and need for further support.

A questionnaire was used in structured interviews with subjects. The questionnaire is shown in the

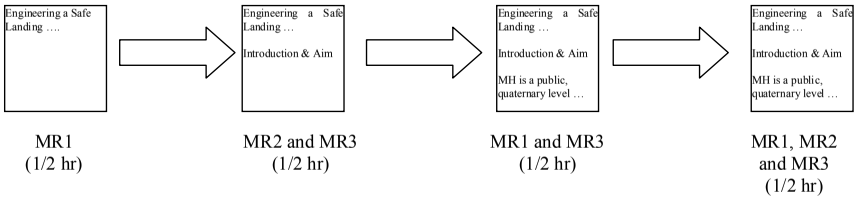



Figure 2: Experimental workflow of four tasks for paper authoring

Appendix. The questionnaire contains open-ended questions where many of the questions address use of:

- workflow only
- collaborative word processor only
- both workflow and collaborative word processor

Open-ended questions are used because they allow the subjects to describe their experience in carrying out the task, and they can capture what actually happens when using the tools. These results reflect how effective or ineffective the tools were in supporting research paper generation, and therefore, generation of many other similar documents types. Quantitative results were not sought because there were three subjects in the experiment. The intention of the study was to provide some initial insights into the use of these technologies in this medical domain. Such insights can guide further research into usage or even development of these tools.

Some questions of the questionnaire can be clarified at this point (see Appendix):

- Question 3 compared the usual situation the researchers experienced when *finalising a paper* where all researchers were "huddled around" the same computer using Microsoft Word with the experimental situation where researchers were seated at their own work computers using CoWord (represented by the *last task* in Figure 2).
- Question 4 sought the researchers' responses to collaborating on the document with another researcher in the experiment (the two central tasks in Figure 2) compared to the previous situation where they worked on their own.
- Question 5a. determined if the subject found using CoWord an overall success. If CoWord was found successful overall, Question 5b. determined what problems and difficulties were experienced, despite overall success. Note that Question 7a and 7b, and 9a and 9b are similar in nature to Question 5a and 5b.

The first author of this paper interviewed each subject individually using structured interviews. The subjects wrote their responses onto their own copies of the questionnaires during the interviews (self-completed questionnaires), and were questioned by the author to clarify any issues related to their responses. The qualitative contents of these questionnaires were analysed to determine the findings presented in the next section.

## 4 Results of Analysis

Questionnaires were analysed by summarising and synthesising the written responses of subjects. The results from the questionnaires are covered in this section. The purpose of this paper is not to cover responses to all questions of the questionnaire, but to focus on major results found.

Firstly, in response to the "Background" question in the questionnaire, the years of experience of subjects in electronic word processing and electronic mail were: 6 years, 8-10 years and 20 years. Hence, it is clear that our subjects had sufficient expertise in the basic tools of email and Microsoft Word.

### 4.1 Question 3 — Finalising the Paper

One subject stated that it was "Effective working on your own computer in contrast to all sitting around a single computer" and that this "increased work speed and efficiency" in writing the document.

Another subject stated that it was better to work in this way because all subjects were able to work together at the same time. However, this subject stated that it was important for all the researchers to have a strategy in order for all of them to work at the same time most effectively. The subject gave the example of the three researchers carrying out possibly different tasks: one researcher works on content of the document, one researcher verifies spelling of the document and the last researcher makes suggestions or comments on the content of the document.

The final subject commented that the finalisation task depicted in Figure 2 was easy to carry out, apart from the software "glitch" in CoWord where there was "jumping" of the page when multiple users worked on the same page. The subjects complained about this problem of "jumping" a few times—a software-specific bug that is not found across collaborative editors and word processors, but happens to exist in the current version of CoWord.

### 4.2 Question 4 — Collaborative Authoring Sessions

One subject's response to this question was that the new way of collaborating simultaneously "worked smoothly and decreased issues around emailing multiple copies of the same document. As all users were working on a single file it is easier to keep track of workflow [the subject means: the flow of work] – and document changes" and was "more time efficient, potentially as do not need to wait for one person to finish for the next to start". The complaint this subject wrote in their response, which did not relate to this question, was that there were "Some limitations in the actual 'word' software: decreased functionality. E.g., lack of copy-and-paste/table options". The subject was indicating that the current version of CoWord has some deficiencies regarding copy-and-pasting and creation of tables in documents. Again, these are difficulties with the current version that CoWord developers need to address, and are not peculiar to all collaborative editors or word processors.

The second subject's response was that such collaborative sessions were effective and able to be

performed very easily. This subject found no difficulties in collaborating with another researcher and that the software did not provide any major barriers to collaboration.

The remaining subject stated that there was "No problem at all doing this phase", and that problems did not occur since the researchers were working on different sections during the sessions.

### 4.3 Question 5a and 5b — General Success and any Difficulties in using the Word Processor

The first subject indicated it was generally successful to use CoWord to author the paper. However, the main problem faced by this subject was that there was "some lack of synchronicity between views seen by multiple users" that "led to some edits being performed multiple times as viewer saw different versions of the document". This response reflects a bug in CoWord. For instance, given the word "Summary", one subject deletes "Su" from the word and another subject adds "Su". The first researcher doesn't see "Summary", but sees "mmary".

The second subject gave his/her own personal rating of 6/10 to reflect the success of CoWord in the authoring process. Overall, the subject was satisfied with using CoWord, and the reasons for giving a lower-than-expected rating include the following examples:

- "insufficient Word functionality, e.g., right click functions, paragraph control/formatting, bullets and numbering"
- "jumpiness [of the page when multiple users worked on the same page]"
- "word chopping" [parts of words being cut as mentioned above by the first subject]

It can be seen that the problems of the current CoWord version have been the major factor in reducing the overall satisfaction that the researchers had in using the tool for collaborative authoring.

The final subject stated that CoWord was "Overall v. [very] successful. It is extremely useful to have one document rather than multiple copies which become v. [very] confusing".

Because all subjects answered Question 5a and 5b, there was no response to Question 6 about lack of success of CoWord.

### 4.4 Question 7a and 7b — General Success and any Difficulties in using Workflow

A response to this question was that "Collaborative workflow has great potential to assist preparing an entire document with multiple authors". This subject was simply indicating that they found the workflow tool was able to achieve the purpose of driving the collaboration over a document by a group of medical researchers.

Another subject stated that they found workflow was very successful in achieving the goal of automating the entire authoring process and gave the tool a personal rating of 8/10. The subject suggested that it would be even better "if users can be automatically notified by email or when they log on that a doc[ument] was ready for them to work on". We had pursued the integration of email with TrackNShare to carry out this email notification before the experiment was conducted, however, we were unable to achieve this.

The final response to this question by a subject was that the software was "simple and to use and there were no problems in saving, forwarding or opening documents".

Because all subjects answered Question 7a and 7b, there was no response to Question 8 about lack of success of CoWord.

### 4.5 Question 9a and 9b — General Success of Workflow and Word Processor

The subjects gave reasonably simple responses to these questions, reflecting that workflow fulfilled its purpose in providing access to the document whilst simplifying the effort of the authors. The first subject stated that the workflow system "works fine", but that the problems experienced were associated with CoWord (e.g., "jumping" and "synchronicity of content"). The second subject remarked that together these tools were effective generally and gave a rating of 6/10 because of the problems with CoWord (e.g., limited functionality of CoWord). The final subject responded that the combination was generally very successful and "would use it again".

Because all subjects answered Question 9a and 9b, there was no response to Question 10 about lack of success of the combination of the tools.

## 5 Discussion

From our results, there was major success in using a workflow system and a collaborative authoring system together to write a very important document in medical research. It can be seen that workflow facilitated distributed, synchronous collaboration (workflow was "making it all happen"). Workflow provided relief in avoiding confusion over multiple copies of the paper. Given the subjects worked only on one version of the paper, workflow also assisted the subjects by preventing them from dealing with the routing of the document, and from being concerned about its storage and versioning.

A representative collaborative authoring system was able to achieve the overall goal of writing the document with a minimum of fuss. In comparison to

the usual way of working where all researchers were positioned at the one, same computer to finalise a paper together, CoWord presented no problem in working simultaneously from separate computers. This new way of working was seen as more efficient as stated by one of the researchers. Of course, this is all contingent upon an agreed group strategy for collaboration. The researchers were pleased to replace working at the same computer using Microsoft Word with working from their own computers using CoWord.

This experiment had been carried out on the basis that there were four tasks involved in authoring a paper. This number of tasks was deliberately fixed in advance; we were more interested in the effectiveness of the tools in carrying out the process rather than completing the entire document using the two tools. In reality, authoring a medical research paper would clearly involve more than four tasks. Indeed, the exact number of tasks a particular team would require would not be known in advance and would be dynamic. Hence, flexible workflow would be relevant to this process. It is future work to investigate usage of flexible workflow for such a process. The idea of the current experiment was to use a simple scenario so that a pre-determined, relatively small number of tasks would be used to focus upon the usefulness and effectiveness of the combination of workflow and collaborative authoring.

## 5 Conclusion

This paper presented an experiment on authoring of a medical research paper using workflow and collaborative authoring systems. Findings of the experiment, based on a questionnaire used in structured interviews, were presented. These two types of tools proved sucessful in assisting researchers to achieve their goal. Some annoyance was caused because of small-scale bugs and unimplemented Microsoft Word functionality in the collaborative authoring system. However, this is not a tremendous problem and requires further effort from the developers of the system.

We are interested in investigating the use of these collaborative tools in other medical document processes and determining how effective they are in supporting such processes. Collaboration is an important component in medical applications, not only between researchers, but between doctors, patients and administrators. Hence, investigating the usefulness of collaborative tools in medicine is an interesting avenue to pursue.

## References


[1] Workflow Management Coalition. Workflow Management Coalition Terminology & Glossary. Retrieved 31 May 2005, from http://www.wfmc.org/standards/docs/TC-1011_term_glossary_v3.pdf.

[2] K. Ho, H. Leong and W. Lam. A collaborative word processing systems using a CORBA-based workflow system. In *3rd International Symposium on Distributed Objects & Applications*, pages 176-185, Rome, Italy, 17-20 September 2001.

[3] S. Xia, D. Sun, C. Sun, D. Chen and H. Shen. Leveraging single-user applications for multi-user collaboration: the CoWord approach. In *CSCW '04*, pages 162 – 171, Chicago, USA, 6-10 November 2004.

[4] Memetex, Inc. TrackNShare Real Time Business Process Manager. Retrieved 28 July 2005, from http://www.memetex.net.

[5] IBM. Lotus Software. Retrieved 2 June 2005, from http://www-306.ibm.com/software/lotus/sw-atoz/index.html.

[6] C. Ames, S. Burleigh and S. Mitchell S. WWWorkflow: World wide workflow. In *HICSS 1997*, pages 397-404, Maui, Hawaii, 7-10 January 1997.

[7] G. Fakas and B. Karakostas. A peer to peer architecture for dynamic workflow management. *Information and Software Technology*, Volume 46, Number 6, pages 423-431, 2004.

[8] A. Schill and C. Mittasch. A generic workflow environment based on CORBA business objects. In *Middleware'98*, pages 18-34, The Lake District, England, 15-18 September 1998.

[9] G. Vossen and M. Weske. The WASA2 object oriented workflow management system. In *ACM SIGMOD International Conference on Management of Data*, pages 587-589, Philadelphia, USA, 1-3 June 1999.

[10] X. Yongyi and Z. Weishi. Component-based workflow architecture of a distributed software process management system. In *3rd International Conference on Quality Software*, pages 204-210, Dallas, USA, 6-7 November 2003.

[11] FINEOS Corp. Ltd. Technology – Workflow manager. Retrieved 2 June 2005, from http://www.fineos.com/technology/workflow_manager/index.htm.

[12] A. Abrahams, D. Eyers and J. Bacon. An asynchronous rule-based approach for business process automation using obligations. In *2002 ACM SIGPLAN Workshop on Rule-based Programming,* pages 93-103, Pittsburgh, Pennsylvania, 5 October 2002.

[13] L. Hassell and J. Holmes. Modelling the workflow of prescription writing. In *2003 ACM*

*Symposium on Applied Computing*, pages 235-239, USA, 9-12 March 2003.

[14] R. Baecker, D. Nastos, I. Posner and K. Mawby. The user-centred iterative design of collaborative writing software, In *InterCHI'93*, pages 399-405, Amsterdam, 24 - 29 April 1993.

[15] J. Begole, C. Struble, C. Shaffer and R. Smith. Transparent sharing of Java applets: a replicated approach. In *10th Annual ACM symposium on User Interface Software and Technology,* pages 55 - 64, Alberta, Canada, 14 -17 October 1997.

[16] T. Dobrowolski. MoonEdit. Retrieved 3 June 2005, from http://moonedit.com/

[17] Y. Yang, C. Sun, Y. Zhang and X. Jia. Real-time cooperative editing on the Internet. *IEEE Internet Computing*, Volume 4, Number 1, pages 18-25, 2000.

[18] M. Locasto, M. Hulme, R. Gladysiewicz, J. Tracy and U. Wolz. Clay: synchronous collaborative interactive environment. *The Journal of Computing in Small Colleges,* Volume 17, Number 6, pages 278-281, 2002.

[19] C. Qingzhang, H. Zangyin and Y. Kezhen. XML-based collaborative documents model design, In *8th International Conference on Computer Supported Cooperative Work in Design*, pages 24-28, Volume 1, Xiamen, China, 2003.

[20] M. Weber, G. Partsch, S. Hock, G. Schneider, A. Scheller-Houy and J. Schweitzer. Integrating synchronous multimedia collaboration into workflow management. In *International ACM SIGGROUP Conference on Supporting Group Work*, pages 281-290, Phoenix, Arizona, 16-19 November 1997.

[21] T. Hodel, H. Gall and K. Dittrich. Dynamic collaborative business processes within documents. In *SIGDOC 2004*, pages 97-103, Memphis, USA, 10 – 13 October 2004.

[22] G. Joeris. Cooperative and integrated workflow and document management for engineering applications. In *Eighth International Workshop on Database and Expert Systems Applications*, pages 68-73, Toulouse, France, 1-2 September 1997.

# Appendix

## *Questionnaire*

### Background

How long have you been authoring documents using e-mail and Microsoft Word?

### Experiment

**1.** How successful/unsuccessful did you find editing the document when other users are also *editing at the same time*?

**2.** How did you find forwarding the documents using workflow instead of e-mail?

**3.** Scenario 1: Editing a document with Microsoft Word, all users sitting at one computer.

At the final phase, you finalize a document at the same time while sitting at your own computer. How did you find finalizing a document at this phase using the collaborative word processor sitting at your own computer when compared to Scenario 1?

**4.** Scenario 2: Editing the document all alone with Microsoft Word, and then forwarding to the next user.

Before the last phase of finalization of documents, there were two phases of authoring the documents sitting at your own computer. How did you find two of you editing the documents using the collaborative word processor during these phases when compared to Scenario 2?

Collaborative word processor:

**5.**

**a.** How generally successful was it editing the documents using a Collaborative word Processor?

**b.** If there were any problems/difficulties, what are they?

**6.** If using the collaborative word processor was not generally successful what were the reasons for this?

Workflow:

**7.**

**a.** How generally successful was workflow software in assisting users to carry out the entire process of authoring a document?

**b.** If there were any problems/difficulties, what are they?

**8.** If using workflow was not generally successful what were the reasons for this?

Workflow and collaborative word processor:

**9.**

**a.** How generally successful was it authoring documents using a collaborative word processor and workflow?

**b.** If there were any problems/difficulties, what are they?

**10.** If using collaborative word processor and workflow was not generally successful what were the reasons for this?

**11.** What suggestions do you have to improve the workflow and collaborative editing?

# Applying Formal Concept Analysis to Semantic File Systems Leveraging Wordnet

*Ben Martin*

Information Technology and Electrical Engineering
The University of Queensland
St. Lucia QLD 4072, Australia
*monkeyiq@users.sourceforge.net*

*Peter Eklund*

School of Economics and Information Systems
The University of Wollongong
Northfields Avenue, Wollongong, NSW 2522, Australia
*peklund@uow.edu.au*

**Abstract** *Formal Concept Analysis can be used to obtain both a natural clustering of documents along with a partial ordering over those clusters. The application of Formal Concept Analysis requires input to be in the form of a binary relation between two sets. This paper investigates how a semantic filesystem can be used to generate such binary relations. The manner in which the binary relation is generated impacts how useful the result of Formal Concept Analysis will be for navigating one's filesystem.*

**Keywords** Document Databases, Document Management

## 1 Background

Semantic File Systems convey the idea that a document can be found in variety of ways according to its content and the search requirements of the user.

A Semantic File System (SFS) unifies data sources through an extended File System interface. Two of the core features of an SFS are the extraction of metadata from files and the ability to create virtual directories showing filesystem objects which satisfy a query defined on the extracted metadata [10]. For an SFS, metadata describes the content of a filesystem object, for example, the width and height of an image. Metadata for files is stored (or derived) for each file and presented though an Extended Attribute (EA) interface [12, 11].

We focus on a particular opensource SFS implementation: libferris [12, 1]. Motivation for the use of Formal Concept Analysis on filesystems includes the ability for the system to handle over specified queries, the provision of an ordered grouping on query results and the ability to switch between query and navigation [12, 7].

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

Formal Concept Analysis [9] takes as input a binary relation between two sets and generates as output a set of "formal concepts" and an ordering relation over them. The formal concepts are a binary set of maximal clusters based on objects (files) and file attributes. An order relation is induced over the formal concepts and is referred to as a concept lattice. The input binary relation is referred to as a formal context. A formal concept (or simply concept), can be thought of as the largest connection between two sets which contains a specific element of one of the sets. Typically the two sets which the binary relation is held on are referred to as the Objects $G$ and Attributes $M$. Thus for each object $g \in G$ one could consider a concept $(A \subseteq G, B \subseteq M)$ to be generated. It is natural for many objects to generate the same formal concept and hence the technique is a form of unsupervised machine learning.

It is natural for one to consider the files and directories of a Semantic File System as forming the object set for Formal Concept Analysis. In this way one might wish to use the metadata for his/her files to form the attribute set for Formal Concept Analysis. If one has some binary metadata about a file, for example is-character-device, then its presence can be taken to directly imply a connection in the input binary relation $I \subseteq G \times M$ for Formal Concept Analysis.

The requirement for input to Formal Concept Analysis to be in the form of a binary relation presents challenges when applying it to an SFS in general. This is due to the fact that the metadata attached to the files in an SFS are rarely simple binary values. Also some metadata which at first appears to be binary may have additional structure which should be taken into consideration. For example, the libferris Semantic File System has the notion of emblems. An emblem allows the user to categorize their files either explicitly or implicitly [11]. This may at first appear to be a simple binary attachment where for a specific emblem a file either has that emblem associated or it does not. However the em-

blems in libferris themselves form a partially ordered set and as such the association of an emblem $x$ also conveys information about a files' association with all the parent emblems of $x$ in this partial order.

Various solutions have been proposed in the Formal Concept Analysis community to allow its application on non binary input. The input to this process is called a many-valued context $(G_y, M_y, W_y, I_y)$ and the process involves taking values, $W_y$, the many-valued input for each attribute $M_y$ has, into consideration to generate binary attributes as output. These solutions include conceptual scaling [9] and logical scaling [14].

Some standard scaling techniques include: nominal, ordinal and interordinal. A nominal scale for an attribute $M_y$ generates a new attribute in the output for each value of $W_y$ which $M_y$ takes in the input. If an object $g \in G_y$ has value $w \in W_y$ for attribute $m \in M_y$ then it will have attribute $M_{my}$ in the output. An ordinal scale takes an attribute $M_y$ which has a naturally ordered set of values $W_y$ and divides the input value range into many linear intervals to form output attributes. An interordinal scale combines two ordinal scales, one using $\leq$ the other $\geq$ on its ordinal range.

## 2 Introduction

The libferris Semantic File System includes extensive indexing support for the storage of EA for files and the application of Formal Concept Analysis to this index. This paper explores how the many-valued data that a Semantic File System contains [3] can be transformed into a binary relation suitable for the application of Formal Concept Analysis.

A simplified overview of the process of applying Formal Concept Analysis to the Semantic File System is now described. Firstly, the filesystem is indexed by libferris for fast retrieval. We shall refer to this index as an "findex" to separate it from the other uses of the word index. Various clients, specifically designed for Formal Concept Analysis, are then applied to the findex to generate a concept lattice. The concept lattice is itself stored as part of the findex and to allow for subsequent reexamination. A concept lattice can be represented by a specialized form of Hasse diagram – called a line diagram – though which its partial order can also be exposed as a Semantic File System by libferris.

It has been found that in many cases some preanalysis for a Semantic File System is needed in order to best expose the Semantic File System without generating a cluttered output.

## 3 Application

The standard scale types of Formal Concept Analysis: nominal, ordinal and interordinal are supported with extensions through three client applications described in Section 3.3. Using a file's URL as metadata to generate formal attributes in various ways is supported as described in Section 3.5. Together with the applications described below there is a method of restricting which documents from an findex are potentially useful. This allows areas of the findex to be negated from query results *en masse*. For example, one might consider only documents under `/usr/local` to be of interest for a particular analysis and so restrict all results to also satisfy this condition.

To demonstrate, an findex was created on a Fedora Core 4 Linux machine using libferris 1.1.54 of 201,759 files in `/usr/share/`. All libferris clients for creating input for Formal Concept Analysis use either the `gf-create-fca-scale` or `ferris-create-fca-scale` prefix in their command names. The clients are subsequently referred to without prefix.

Section 3.1 discusses application to nominal binary data, section 3.2 applies to geospatial metadata for files, section 3.1 discusses application to numeric domains. The application to NSA SELinux [13] follows in section 3.4 followed by the use of Wordnet [8] to improve the structure of concept lattices created from file URLs in section 3.5.

### 3.1 Scaling nominal orders

In addition to the standard treatment of nominal scaling [9, 5], two new capabilities for handling ordering over nominal attribute creation have been added.

The first ordering capability is to handle MIME type like strings such as `image/png` by allowing the values of the distribution to be split into distinct parts and have common parent attributes created. Following the MIME example, a common parent for all image files would be the new `mime.image` attribute. Using this form of nominal scaling an ordering can be introduced based on the values of the distribution which will help to generate a taller, narrower concept lattice [5].

The second ordering capability is to take advantage of the ordering over the emblems when performing nominal scaling via an emblem. The ordering on the emblems is a partial order allowing reasonable flexibility in how one designs emblem categories. The ability to handle entire downsets relative to the emblem ordering when generating a formal context allows one to see their lattice including the influence of their emblem ordering. Given an ordered set $P$ and $Q \subseteq P$ then $Q$ is a downset iff $x \in Q, y \in P$ and $y \leq x$ then $y \in Q$.

### 3.2 Scaling Geospatial information

Geospatial metadata is exposed through two cooperating interfaces in libferris. These are the latitude and longitude EA and the emblem system. Geospatial emblems are those which are a child of the `libferris-geospatial` emblem in the emblem partial ordering. Interaction with the filesystem for tagging and retrieval is usually simpler when emblems with city or place names are used instead of world coordinates.

As the emblem system is employed the scaling methods of Section 3.1 are also applicable for geospatial values. A major advantage of the emblem geotagging system coexisting with the latitude and longitude system is the ability to handle geospatial regions. The emblem partial order can be used to define geospatial regions that expand from point locations to physically containing regions. For example, the Sydney Opera House might be given a specific emblem with Sydney as its parent. The Sydney emblem may have Australia which itself has `libferris-geospatial` as its parent. If less specific emblems in the partial order define containing geospatial regions then the downset handling in Section 3.1 can be used to introduce geospatial refinements into the concept lattice.

Without the ability to represent geospatial regions through the emblem partial ordering in this way one would have to explicitly define the boundaries using bounding box constraints on the latitude and longitude for the region. Consider the difficulty in defining the boundary of the city of Sydney using only equality constraints on latitude and longitude.

### 3.3 Scaling numeric ranges

Three commands exist for creating formal contexts from numeric data in libferris. These are `numeric-ordinal`, `numeric` and `gf-numeric`.

The `numeric` client can create many binary attributes each exposing a numeric interval of the input data as is standard for Formal Concept Analysis. For example, consider scaling a numeric range of $\{1, 2, ..., 20\}$ into four attributes at an even interval of $5$ using $\leq$ as is standard in Formal Concept Analysis. This will produce four attributes with higher successive values having less matches due to the $\leq$ relation.

The standard application of ordinal scaling preserves both linearity *and* density for the input [6]. Due to the intermixing of other attributes in a concept lattice it is hard to take advantage of the preservation of density information. When one places these four attributes alongside another ten attributes and generates a concept lattice the relation between $\leq 10$ and $\leq 15$ is not so immediately obvious from the concept lattice. One can see the order of the two attributes but the density information is lost due to the introduction of the other attributes.

For some value distributions using a linear interval for the range is ineffective. For example, if one is to scale the values for the file size metadata then the distribution of values may be very ineffectively presented when split into a low number of linear intervals. To overcome this issue the data-driven-scale option was added to allow numeric distributions to be scaled taking the value distribution into account. This option will make an output which will have smaller intervals where many files have similar values and larger intervals where few files match the interval.

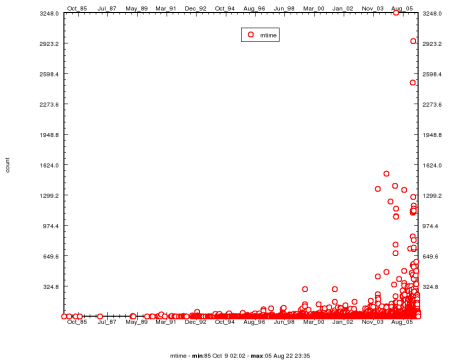



Figure 1: Plot of the modification time of 201,759 files from /usr/share/. Horizontal axis shows time from October 1985 to preent day with almost 2 years between graduations. Vertical axis ranges from 0 to 3248 with around 235 files seperating each graduation.

The std-deviations option has been added to handle simpler value distributions by allowing output to be generated based on the mean and variance of the input distribution.

One can manually select where intervals begin and end using the GTK+ `gf-create` client. Figures 1-4 were generated with `gf-create`. For value distributions which neither fit direct data frequency nor standard deviation models the ability to explicitly choose where intervals begin and end on a value frequency plot can generate a small number of meaningful attributes. For this purpose an interactive graphical client was create allowing intervals to be selected with the mouse.

The plot for the modification time (mtime) of the findex is shown in Figure 1 and the metadata status change time (ctime) in Figure 2. One can see that although modification was more frequent in recent times the ctime plot has explicit natural clusters of values. Such clusters are likely due to large scale system administration activities such as distribution release upgrades. Using the graph a small number of meaningful attributes can be created based on major system update activities.

An EA was added to the libferris system to support the ability for many versions of a file's metadata to exist simultaneously in an findex [2]. This EA returns the current system time when it is read. As expected, the plot for this attribute gives valuable information about when indexing sessions were held as shown in Figure 3. Looking at Figure 3 one would be lead to create three formal attributes, one for each of the major groupings of matching files.

The width EA presents the width in pixels of a file. For many systems this EA must be handled explicitly

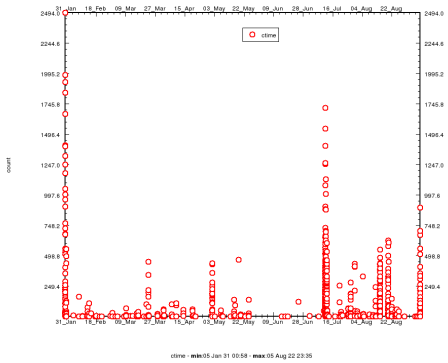



Figure 2: Plot of the ctime of 201,759 files from /usr/share/. The ctime for a file changes whenever any of its metadata (except atime from lstat) changes. Horizontal axis shows time from 31st January to 05 August 2005 with two and a half weeks between graduations. Vertical axis ranges from 0 to 2494 with around 250 files seperating each graduation.

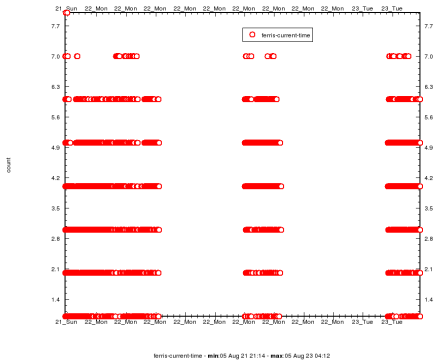



Figure 3: Plot of the ferris-current-time EA of 201,759 files from /usr/share/.

because a small number of extremely large images can easily distort simpler methods of splitting the value distribution. In this case two images stand out, sgvol.png from the kdemultimedia package is 7,140 pixels wide and sunclock_huge_earthmap.jpg from the sunclock_huge_earthmap package is 10,800 pixels wide. All other image files in the findex are below 3,500 pixels wide. The width plot is shown in Figure 4. One can also start from the megapixel count of images as more generalized overview of image size to generate

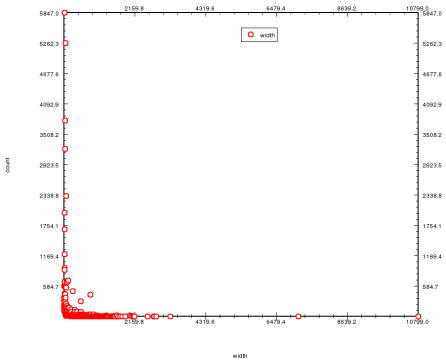



Figure 4: Plot of the the width of image files from /usr/share/.

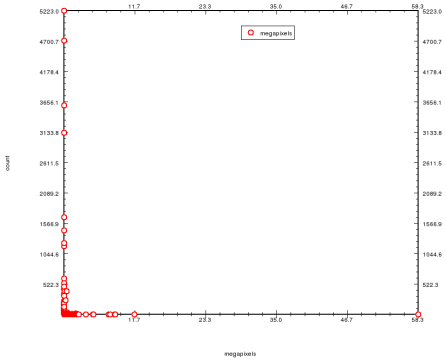



Figure 5: Fewer plot points but a similar overall trend to the width plot. Plot of the the megapixels of image files from /usr/share/.

formal attributes. As can be seen from Figure 5 there is a similar trend as to the width plot.

Two concept lattices were generated using the width EA and the modification time for the examples 201,759 files. Both scale the width and modification time metadata using 7 formal attributes for each. The first one shown in Figure 6 uses the standard linear ranges to generate the formal attributes. Shown in Figure 7 is the concept lattice that results when dividing the input ranges based on value density.

Because the formal attributes in Figure 7 are data driven there is much more interaction between concepts in the resulting concept lattice.

For some numeric EA such as: group-owner-number, user-owner-number the user may wish to

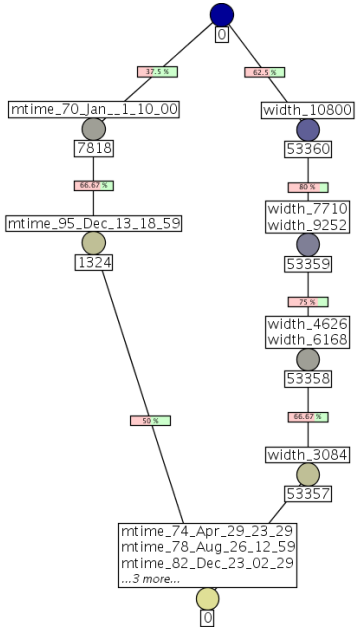



Figure 6: 7 formal attributes for each of mtime (modification time) and width using a standard linear range division. Concepts are represented as circles. Labels above a concept show the formal attribute which is introduced by that concept and labels below a concept show the number of filesystem objects which match that concept or one of its refinements. An introduced formal attribute is a formal attribute for which this concept is the highest one in the lattice with that attribute. Thus, where a concept has an introduced formal attribute all concepts reachable transitively downwards will also have this formal attribute.

explicitly specify the range for each formal attribute based on knowledge of the computer system. For example, on many Linux installations the numeric user and group identifiers above 500 are used for normal user accounts.

### 3.4 SELinux

Security Enhanced Linux (SELinux) [13] allows modern Linux installations to offer Mandatory Access Control (MAC) as well as the more familiar Discretionary Access Control (DAC). Under DAC, file access is granted or denied based on the user running an application. Assume that my user account as read and write access to my thesis and read access to my music collection. Under DAC a music player has the ability to overwrite my thesis just as xemacs can read my music files. With MAC programs can be allowed access only to the files that are required for them to operate. For example, using MAC I can disallow my media player access to any files relating to my thesis. It should be noted that my user account will still be the owner of my music files and thesis though the media player run by me will be disallowed access to some files owned by my user account.

SELinux information which is attached to files is comprised of three datum: the identity, role and type. The identity is a SELinux user account, the role is ignored for files and the type is the primary security attribute for making authorization decisions.

In a minimal installation one has an SELinux user_u account which is shared by all users in a similar category and a system_u for daemon usage. The example 201,759 files have three identities: root, user_u and system_u. Also there are nine types: etc_t, fonts_t, httpd_sys_content_t, lib_t, locale_t, man_t, shlib_t, snmpd_var_lib_t, and usr_t.

A very high level view of how access is granted or denied follows, for details see [13]. Each process also has an associated SELinux context. Access is granted or denied based on the SELinux context of the process and the file together with the operation requested to be performed. As such viewing only the SELinux context for files provides an incomplete picture of overall security policy.

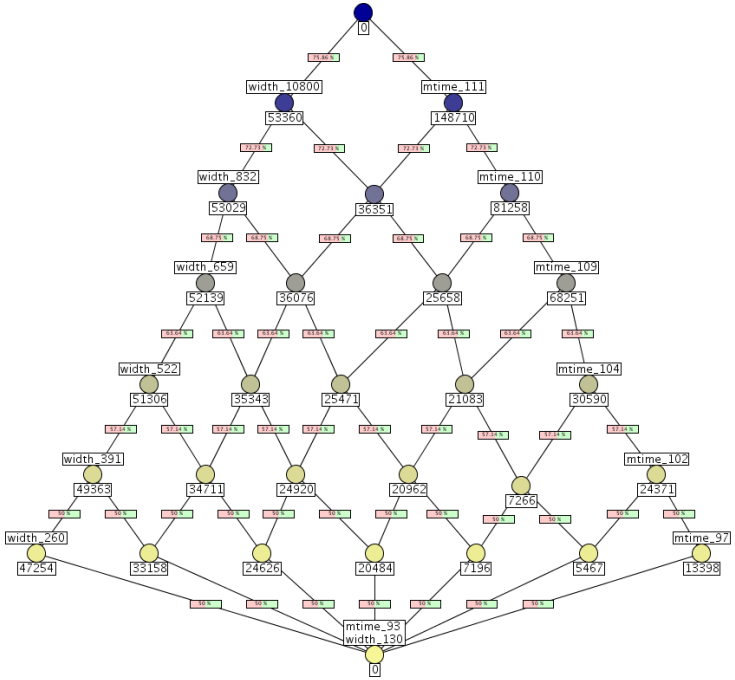



Figure 7: 7 formal attributes for each of mtime (modification time) and width. Formal attributes are generated based on the density of the input metadata.

Using the SELinux type and identity of the example 201,759 files the concept lattice shown in Figure 8 is generated. The concept 11 in the middle of the bottom row shows that user_u identity is only active for 3 fonts_t typed files. Many of the links to the lower concepts are caused by the root and system itentities being mutually exclusive while the system identitiy combines with every attribute that the root identity does.

### 3.5 Structuring with URLs

Often the URL for a file is comprised of metadata forming an ad-hoc hierarchy [4]. To put such metadata into the URL itself requires arbitrary decisions about the ordering of such metadata. For example, one must decide if they are to first classify a file by its conference name or conference year in the URL `.../adcs/2005/martin-eklund/...`.

The `scale-urls` client creates a formal context from the directory components in URLs. Additional processing can be performed to present a more attractive and useful concept lattice. For example, heuristics can be used to strip version information from directory names such as `java-1.5.0`.

Wordnet [8] is also employed to explicitly allow the generation of formal concepts for hypernyms of common directory names. Explicit hypernym concepts are generated as follows: each URL is divided into its directory name components with a number of the rightmost path components being dropped (normally just one, the filename), each directory name is then stripped of version information and added to the set $D$. Many such preprocessed directory names $d \in D$ are then candidates for use with Wordnet. If $d$ can be found in Wordnet then its synonym set $X$ is found and all the hypernyms for $X$ are collected. When two or more $d$ have the same synonym set $X$ then the hypernyms for $X$ are emitted into the formal context with a prefix "wn_" to denote that they have been mechanically added.

The semantic commonalities between directory names are made more explicit in the output concept lattice using the Wordnet hypernym associations. Another advantage is that because the "wn_" attributes effectively form the join of many existing attributes they are closer to the top of the concept lattice. If the concept lattice is being read in the usual way from top to bottom or is itself being navigated as a filesystem the placement towards the top of the lattice advantageous to have these wordnet attributes to assist in navigation to the desired concept.

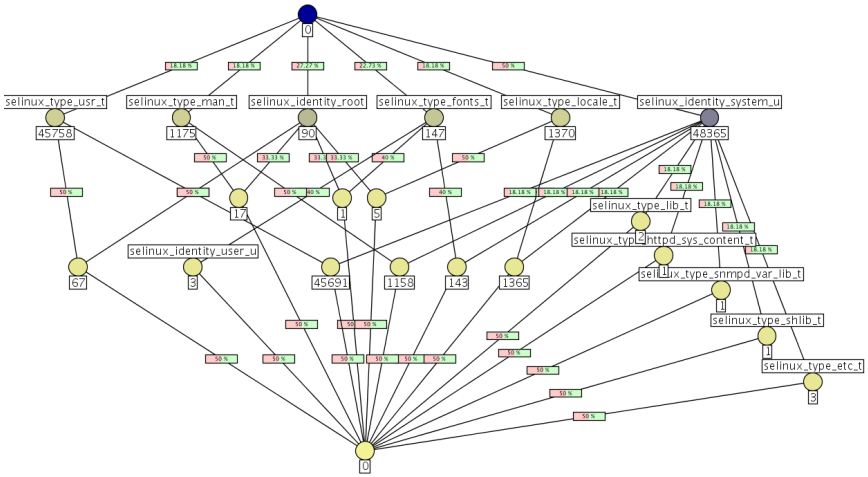



Figure 8: Concept lattice for SELinux type and identity of files in /usr/share/ on a Fedora Core 4 Linux installation. The Hasse diagram is arranged with three major sections; direct parents of the root are in a row across the top, refinements of selinux_identity_system_u are down the right side with combinations of the top row in the middle and left of the diagram.

For the example findex, dropping only the final directory component (i.e., filename) from each URL the wordnet scheme above generated 403 new "wn_" attributes. An example of one such attribute is wn_article which is equal to feature $\vee$ paper.

Shown in Figure 9 is small example concept lattice where the two files of interest are both masked from the top concept by attributes which one does not immediately consider related to the child concept. Consider that both the concepts labeled with $a1$ and $a2$ may have many more child concepts than those shown. The introduction of a semantically more general attribute form wordnet in Figure 10 may allow the user greater ease in locating their desired concept.

## 4 Conclusion and Future Directions

The application of Formal Concept Analysis to a Semantic File System is not straightforward. One must choose what metadata to use to construct the input for Formal Concept Analysis: the formal attributes. This stage is critical to the generation of an interesting concept lattice.

This paper has shown some methods of applying Formal Concept Analysis to an Semantic File System. The generation of formal attributes using Wordnet can have a positive effect on the usability of a concept lattice when using URLs. This is very important as the file paths are normally the only metadata explicitly supplied by the computer user.

When considering ordinal metadata the density of files with particular values of metadata can be taken advantage of to generate fewer formal attributes for the same metadata. When metadata which is ordinal is to be transformed into formal attributes one can have fewer attributes if they use non linear ranges to break up the metadata. The density information that is retained by using linear ranges to break up ordinal metadata is not readily apparent in a concept lattice with many formal attributes from different metadata sources.

## References

[1] libferris, http://witme.sourceforge.net/libferris.web/. Visited Oct 03.

[2] libferris in the linux journal, http://www.linuxjournal.com/article/7771. Visited Jan 05.

[3] A virtual filesystem on steroids: Mount anything, index and search it. To appear: Linux Kongress 2005 proceedings.

[4] Peter Becker and Richard Cole. Querying and analysing document collections with formal concept analysis. In *Australian Document Computing Symposium (ADCS03)*. University of Queensland, 2003.

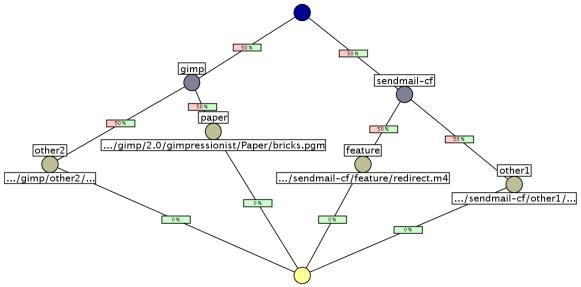



Figure 9: Example lattice with no wordnet augmentation.

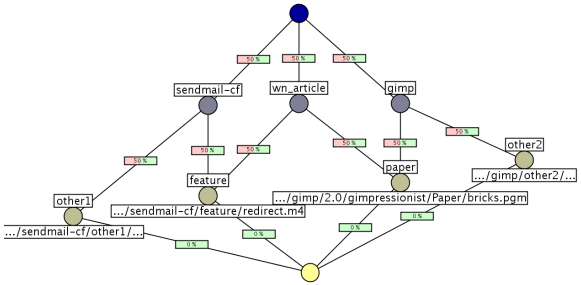



Figure 10: Example lattice using wordnet augmentation, notice how the wn_article concept is the common parent of both feature and paper and is also closer to the top of the lattice than either hyponym.


[5] Claudio Carpineto and Giovanni Romano. *Concept Data Analysis*. Wiley, England, 2004.

[6] Robert Colomb. *Information spaces : the architecture of cyberspace*. Springer, London, 2002.

[7] Sebastien Ferré and Olivier Ridoux. A file system based on concept analysis. In *Computational Logic*, pages 1033–1047, 2000.

[8] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller. Introduction to wordnet: An on-line lexical database. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 112–119, 1990.

[9] Bernhard Ganter and Rudolf Wille. *Formal Concept Analysis — Mathematical Foundations*. Springer–Verlag, Berlin Heidelberg, 1999.

[10] David K. Gifford, Pierre Jouvelot, Mark A. Sheldon and James W. Jr O'Toole. Semantic file systems. In *Proceedings of 13th ACM Symposium on Operating Systems Principles*, ACM SIGOPS, pages 16–25, 1991.

[11] Ben Martin. File system wide file classification with agents. In *Australian Document Computing Symposium (ADCS03)*. University of Queensland, 2003.

[12] Ben Martin. Formal concept analysis and semantic file systems. In Peter W. Eklund (editor), *Concept Lattices, Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, Proceedings*, Volume 2961 of *Lecture Notes in Computer Science*, pages 88–95. Springer, 2004.

[13] Bill McCarty. *SELinux: NSA's Open Source Security Enhanced Linux*. O'Reilly & Associates, Sebastopol, California, 2004.

[14] Susanne Prediger. Logical scaling in formal concept analysis. In *International Conference on Conceptual Structures*, pages 332–341, 1997.

# Biomedical Named Entity Recognition System

**Jon Patrick and Yefeng Wang**
**Sydney Language Technology Research Group**
**School of Information Technologies**
**University of Sydney**
{jonpat, ywang1@it.usyd.edu.au}

**Abstract** *We propose a machine learning approach, using a Maximum Entropy (ME) model to construct a Named Entity Recognition (NER) classifier to retrieve biomedical names from texts. In experiments, we utilize a blend of various linguistic features incorporated into the ME model to assign class labels and location within an entity sequence, and a post-processing strategy for corrections to sequences of tags to produce a state of the art solution. The experimental results on the GENIA corpus achieved an F-score of 68.2% for semantic classification of 23 categories and achieved F-score of 78.1% on identification.*

**Keywords** Named Entity Recognition, ME model, Information Retrieval.

## 1 Introduction

The discovery of the human gene and rapid developments in the biomedical domain has produced large amounts of genetic data. This has resulted in exponential growth of biomedical literature over the past few years. MEDLINE, the primary research database serving the biomedical community, currently contains over 14 million abstracts, with 60,000 new abstracts appearing each month. This growth of biomedical literature has given rise to a pressing need for automatic information extraction from the data bank.

Biomedical literature contains a rich set of biomedical entities providing key information to access the knowledge. A biomedical named entity is a word or sequence of words that can be classified as a name or biomedical term, such as protein, DNA, RNA, etc. Named Entity Recognition is the task of identifying and semantically classifying named entities in text. In the biomedical domain, the goal of the biomedical named entity recognition (BioNER) task is to find the biomedical terms such as names of genes, proteins, gene products, organisms, drugs, chemical compounds etc. in texts and classify them into their correct categories. It is a critical step for future automatic processing of biomedical literature to be mounted on a large scale, and further to perform high level biomedical information extraction task such as analysis and question answering.

BioNER consists of two tasks, term identification and term classification. Identification finds the region of a named entity in a text. Its main goal is to differentiate between terms and non-terms without looking at the semantic meaning of a term. However term classification determines the semantic concept of that named entity and assigns it to a biomedical class, such as genes, proteins or DNA.

The named entity recognition in the newswire domain has been studied for a long time and has achieved 90% accuracy [11]. However, named entity recognition in biomedical domain has different characteristics, with an accuracy of only around 70%. Despite the "near human" performance of named entity recognition in newswire domain, many similar strategies do not work well when adapted into the biomedical domain because of the distinctive nature of this task Hirschman et al., [5] Tuason et al.,[15] Shen et al.,[9] Lin et al., [8] and Lee et al.,[6].

First, biomedical named entities are not conventional proper nouns. They are usually unknown words containing uncommon orthographic features such as hyphens, digits, letters, and Greek letters. Furthermore, there are no conventional rules for biomedical term formation.

Second, biomedical terms may have a number of spelling variations. For example, the term *Alpha UF1 cells* may have spelling variations: *Alpha UF-1 cells*, *Uf-1 Alpha cell.* Such variations always cause recognition ambiguity.

Third, ambiguity and inconsistency are often encountered in named entity classification. Many named entities with the same orthographical features may fall into different categories, for example, nested entities of one category may contain an NE of another category, or a NE is composed of two NEs from different categories.

Fourth, complex naming and abbreviation

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

conventions can differ from organism to organism, and class to class. Abbreviations tend to be a short form and coincide with English words such as “can”, “dot”. In addition, the abbreviations are intrinsically degenerate forms, so that one abbreviation can have a number of meanings, depending on the document domain.

Fifth, new named entities are introduced daily as new substances are discovered and some existing terms might change as our understanding changes. The system must be able to recognize new names and unseen names, and this causes difficulties in rule based systems.

In this paper, we explore machine learning (ML) and natural language processing (NLP) techniques to recognize biomedical named entities in text. We present a strategy that is different to previous work on two bases, firstly we use a framework that incorporates as many useful linguistic features as possible for this recognition task, and secondly we use a Maximum Entropy (ME) model as the basis of our machine learning system, finally we apply rule-based post-processing on the classification results.

## 2 Related work

Named Entity Recognition in the biomedical domain is more difficult than in a newswire domain because of the complex name formation of the biomedical terms and our current lack of experience in understanding optimal strategies to solve this task. Current NER approaches include: dictionary based, rule based, machine learning based, and hybrid approach. Due to the spelling variation and complex naming convention of biomedical terms, NER systems that rely on dictionary resources and pre-built rules do not seem to perform well, especially for large scale tasks.

### 2.1 Dictionary and rule based approaches

Early approaches in biomedical named entity recognition typically were dictionary-based approaches and rule based approaches. These approaches use domain specific heuristic rules and rely heavily on existing dictionaries, representative research includes Krauthammer et al.,[19]; Hirschman et al., [5]; Tuason [15]. However, the dictionary-based approaches typically perform quite poorly, with coverage generally only in the range of 10-30%, even allowing for some variability in the form of names. The rule-based systems perform well for existing named entities, but they usually perform poorly on new named entities and it is costly to adapt them to new entity classes. Once a new class is introduced, a set of new rules has to be generated manually. Since there is no standard biomedical term naming convention, the rule building process becomes more difficult as the number of class increases. Furthermore, the rule-based system performs poorly on larger copra, Gaizauskas et al., [4] and Fukuda et al., [3].

### 2.2 Machine-learning approaches

The major problem in machine learning based NER systems is the lack of training data. Before the GENIA corpus 3.0, Kim et al., [20] there was no consistent annotated corpus, so researchers used some small-scale data sets, such as GENIA 1.1 and Bio1. The development of GENIA 3.0 which contains 2000 abstracts provides a standard evaluation data set for the machine learning approach. Many other corpora that derived from the GENIA corpus have been constructed, such as the BioNLP/NLPBA corpus.

The typical machine learning algorithms include Naive Bayes (NB), Support Vector Machine (SVM), Hidden Markov Model (HMM), Maximum Entropy (ME) models, and Conditional Random Fields (CRF). Kazama et al. [6] used an SVM to achieve an F-score of 54.4 on GENIA 1.1. Nobata and Collier [1] incorporated rich features into a hidden Markov Model and achieved an F-score of 75.9 on a primary version of GENIA, which contains 100 medical abstracts, Shen and colleague [9] further enhanced the HMM model by exploring some special phenomena and a rule based postprocessor. They have achieved performance of 66.5 on the GENIA 3.0 corpus. Lin and colleagues [8] adapted a maximum entropy model for biomedical named entity recognition with a post processor, and achieved the performance of an F-score of 72.1. Finally CRF have been introduced into this field. Settles [10], Tsai et al.,[13] and shown good results. (69.9% and 69.8%) on JNLPBA corpus.

A large body of post processing has been proposed for biomedical named-entity recognition, typical work includes Shen et al., [9], Lin et al., [8], Zhou et al., [17]. Shen et al. proposed a rule based system for cascaded named entity resolution. They automatically extract rules from the training corpus. Lin et al., make use of a rule based boundary extension strategy combined with dictionary lookup for reclassification, and this post-processing effectively increases performance by about 20%.

## 3 Modelling the data

### 3.1 GENIA corpus

The GENIA corpus is an annotated corpus of paper abstracts extracted from the MEDLINE database using the MeSH query, *human*, *blood cell* and *transcription factor*. In the current version 3.02, 2000 abstracts are annotated by domain experts with entity tags. The annotation of the biomedical terms is based on the GENIA ontology. The GENIA ontology is a taxonomy of 48 biologically relevant categories. In our system, we recognize 23 distinct entity classes, including Protein, OtherName, DNA, CellType,

CellLine, OtherOrganicCompound, Lipid, MultiCell, Virus, RNA, Tissue, CellComponent, Peptide, BodyPart, AminoAcidMonomer, OtherArtificial-Source, Polynucleotide, MonoCell, Atom, Inorganic, Nucleotide, and Carbohydrate.

## 3.2 Maximum entropy machine learner

In our experiments, we adapt Zhang Le's Maximum Entropy Tool Kit[1]. This is a C++ implementation of OpenMaxent, which contains Generalized Iterative Scaling (GIS) parameter estimation and Gaussian Prior Smoothing algorithms.

One advantage of using a maximum entropy model is that the features need not be statistically independent, and therefore it is easy to incorporate features with dependencies. Some of the features used in this system are strongly dependent, and yet they do not bias the ME model overly much, thus the ME models can yield better probability estimates compared with some other probability based machine learners, such as Hidden Markov Model (HMM) and Naïve Bayes classifier. Another advantage of using ME model is that it is scalable and does not suffer from the data sparseness problem. The training speed of ME model is faster than SVM. Although the training time is a one-time cost in a real word application, however, in prototyping a system, training must be fast enough to allow experimentation with various configurations.

We use a simple bag of word model to represent the language model of the texts hence each token's features are represented by a binary attribute value.

We employ the simplest BIO representation which is widely used in named entity recognition tasks, for example, Kazama et al. [6]. B means the token is at beginning of an NE, I means the token is in an NE, and O means the token is not in a named entity. For each category C, we have B_C and I_C tags to represent the beginning and inside of an NE of that category.

## 4 Feature set

Machine Learning systems typically represent data in terms of a set of features. It is intended that these feature sets encode the most significant aspects of the data for the learning task. In this section, we describe the features we used in our system.

## 4.1 Orthographical features

Orthographical features are used to capture the rendition of words, such as capitalization, digitalisation and punctuation. Orthographic features allow strings to be compared based on their spelling characteristics and are widely used in the biomedical and newswire domains, such as Shen et al., Collier et al., and Tsai et al.,[9,2,13].

Table 1 presents some orthographic feature used in our system. The feature such as AllCaps, for words with only capital letters, is useful to identify biomedical abbreviations. The CapsAndDigits feature is a very strong indicator of entities from Protein, DNA and Othername classes. The comma, colon, bracket, full stop and stop words are useful for detecting the boundaries of named entities. The Greek letters and Roman numerals are often used in biomedical terms, and the feature of LowercaseOnly strongly indicates the non-entity class.

| Features | Example |
|---|---|
| AllCap | ALAS, HIV, RIP |
| SingleCap | B, M, T |
| DigitNumbers | 7, 8 , 41 |
| CapsAndDigit | CD4, MEK1 |
| InitCapDigit | Am80 |
| TwoCaps | FcR, FasL |
| InitCapsLowcase | Ras, Crkl, Ctx, |
| InitCaps | FURa |
| LowCapsMix | dNTPs, dPRL, |
| LowcaseOnly | protein, cell |
| LetterAndDigit | ETh1, h1RaK |
| InitDigit | 15B7, 17q, 1A9 |
| Backslash | / |
| Parenthesis | [,], (, ) |
| Punctuations | ;,:,,,. |
| Hypen | - |
| RomanNumeral | I, II, III |
| HasHyphen | -induced, Eth-1 |
| GreekLetters | Alpha, kappa |
| Other | Other symbols |

Table 1. Orthographic Features with examples.

## 4.2 Part of speech feature

In the newswire domain, the POS features have been shown to be of limited use because the POS features may adversely interact with the use of some important capitalization information [14]. However, POS features are widely used in the biomedical domain [9,3,13,16], because many biomedical entities are in lowercase, and capitalization information in the biomedical domain is not as evidential as that in the newswire domain. Moreover, since the biomedical named entities have many elements, identifying the boundaries is a more difficult task. The POS tagging can help to determine the boundaries, as for example, verbs and prepositions usually indicate a boundary.

[1] http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html

In our experiments, each word is assigned a POS tag feature. The GENIA POS tagger Tsuruoka et al. [14] was used to provide the POS information. The GENIA POS tagger is specifically tuned for biomedical text such as the MEDLINE abstracts, which reported 98.20% accuracy on the GENIA corpus.

## 4.3 Affix features

The prefix and suffix can provide good clues for classifying named entities, and has been widely used in Kazama et al.,[6] Zhou et al.,[17] Tsai et al.,[13] and Lee et al.,[6]. Kazama et al. collected the 10,000 most frequent prefixes/suffixes from the training data while Zhou et al. construct a prefix/suffix list using a statistical method and grouped the prefix/suffix into 23 categories using a weighted score according to the prefix/suffix distributions.

We extracted the affixes from each class of the corpus by their diversity and frequency. Frequent and diverse affixes may have higher priority to be extracted, for example, the suffix *~cyte* is usually a cell type and the suffix *~lipid* is usually a lipid. However, short affixes always conflict with common English words, for example, the suffix *~ase* conflicts with English word "disease". Some common affixes have high diversity and frequency in both entity and non-entity classes, so they do not contribute to the classification, for example, the suffix *~tion*.

We extracted the 3500 most frequent prefixes and suffixes from the training data, we filter the prefixes and suffixes if the root-diversity is less than 5 (examples in Table 2).

| Suffix | Class | Example |
|---|---|---|
| *~nase* | Protein | Kinase |
| *~hift* | Othername | Shift |
| *~esis* | Othername | embryogenesis |
| *~ytes* | CellType | leukocytes |
| *~ycin* | OtherOrganicComp. | rapamycin |
| *~eria* | MonoCell | Bacteria |
| *STAT~* | Protein | STAT1s |
| *NFAT~* | Protein | NFAT2 |
| *path~* | Othername | Pathogenic |

Table 2. Examples of Suffix Features extracted from the corpus.

## 4.4 Unigram named entity feature

The unigram term is similar to the core-term proposed by Fukuda et al., [3] and the single term list in Lee's system [6]. It is a list consisting of all single word named entities, such as *IL-2*, *NF-kappaB*. These terms usually have special surface clues, and appear at the leftmost part of an NE. They can be combined with a head noun to form a new named entity. We extract all unigram named entities from the training corpus, and remove them if their frequency is less than 5.

## 4.5 Head noun feature

The head noun is usually the major element of a noun phrase, which describes the function or the property of the named entity. For example, the *NF-kappaB activation* is the head noun for the named entity *CoCl2-induced NF-kappaB activation*. Some previous works Nobata et al., [18] and Shen et al., [9] show that the head nouns in biomedical named entities can provide significant clues for distinguishing the entity classes. For example, the term *IL-6 kappa B binding factor* is classified as a Protein, and the *L-6 kappa B motif* is classified as DNA. Hence, the classification is determined by the head nouns *binding factor* and *motif*.

We constructed a head noun list by first looking at the rightmost word in a named entity, since the head nouns usually are the last noun in the named entities. A list of head noun candidates was extracted from each named entity class and ranked by frequency, because the most frequent nouns can be a good predictor for that class. We filter out the head nouns with frequency less than 5. Table 3 lists some head nouns extracted from the training data.

| Class | Head nouns |
|---|---|
| Protein | factor, protein, receptor, complex, heterodimer, subunit, kinases, calcineurin, selectin, antibody |
| Other Name | expression, activity, activation, differentiation, apoptosis, phosphorylation, production, assays, levels |
| DNA | promoter, site, gene, element, chromosome, plasmid, repeat, construct, locus |
| Cell Type | Lymphocyte, monocyte, macrophage, neutrophils |

Table 3. Examples of Head Noun Features

## 4.6 Bi-gram phrase feature

We extracted all bi-gram noun phrases from the entities from the training corpus as a feature. We filter the low frequency bi-gram phrases, as we found they cause some negative effects. The bi-gram phrase is similar to the bi-gram head nouns, except we included some high frequency word bi-grams and bi-gram

| | |
|---|---|
| T cell | transcription factor |
| gene expression | cell line |
| virus type | human monocytes |
| signal transduction | Epstein-Barr virus |

Table 4. Examples of bi-gram phrase features.

named entities. Table 4 lists some high frequency bi-gram phrases.

## 4.7 Contextual information

The contextual information is important for this task. The words preceding and following the target words are also used as features in our experiments.

# 5 Post-processing

## 5.1 Fixing inconsistent tag sequence

As we used the B, I, O notation to indicate the location of the token within the NE, the system may produce an inconsistent class sequence such as "*O B_Protein I_DNA O*". However, only a consistent sequence of tags is annotated as a named entity. We identified four types of such inconsistency in classifications, and describe rules using regular expressions to fix these mistakes.

1. I tag without preceding B tags. These tags are mainly due to false positives and partially identified terms. Some lower case words that have been seen in the named entities are classified as *I_OtherName*. We change this type of invalid I tags into O tags as we assumed that fixing these I tags can increase recall to a certain degree. Further inconsistencies of a sequence of I tags is altered so that the first is a B tag.

2. Missing middle I tag. Some middle I tags are classified as O tags, such as "and", "or". We change these O tags according to the preceding B class or I class tag.

3. Inconsistent I tag sequence. The I tag sequence in some long entities may be mixed with I tags from another class, for example, "*O B_DNA I_Protein I_DNA O*". We fix this mistake by changing the inconsistent I tag class into the B tag class.

4. Inconsistent tag sequence due to nested named entity. We found in our experiments, many entities are tagged as "*O B_C1 I_C1 I_C2 O*", where *I_C1* and *I_C2* are tags from two different categories. The I_C2 is usually a head noun and "*B_C1 I_C1*" is a NE from C1. We fix this inconsistency by first checking if I_C2 is in the head noun list, and then assign the NE a class according the head noun's category.

## 5.2 Rule based boundary correction

We found a number of partially identified named entities are due to missing the rightmost head nouns or the leftmost adjectives. We built from the training data a list of head nouns and a list of modifiers that frequently appear in the boundaries of a NE. Then we designed two simple rules to perform the boundary correction which is similar to the boundary extension in Lin et al. [8].

1. NE := NE + headnoun
2. NE := modifier + NE

After the entity recognition is completed by our ME-model and keeping the tag sequences fixed, we applied these two rules on recognized named entities to expand the boundary to the right and left.

# 6 Experiments and discussion

To conduct experiments, we divided the 2000 abstracts into a training set and test set. The training set consisted of 1800 abstracts and the test set consisted of 200 abstracts. The performance was measured by precision, recall and F-score, which are the standard measures for named entity recognition. The accuracy is measured by the number of correctly recognized named entities.

The main computational cost of the ME model is the GIS parameter estimation, which involves computation of each observed expectation, and re-computation of the model's expectation on each iteration. The greater the number of iterations the better the training accuracy. Since the number of iterations we need for the model to converge to an optimal solution is unknown, we ran 2000 iterations for each experiment. The experiment settings are shown in Table 5

| **Training** (#words) | **Testing** (#words) | **Context** (position) | **Iteration** |
|---|---|---|---|
| 415,761 | 43,597 | -2,-1,0,1,2 | 2,000 |

Table 5. Experiment configuration.

## 6.1 The contribution of features

The task was to investigate the contribution of linguistic features to predicting the correct class boundaries and labels. Several experiments were performed using different combinations of features. (Results in Table 6)

The orthographic features (O) are only

| | **Feature** | **P** | **R** | **F** | **Effect** |
|---|---|---|---|---|---|
| 1 | O | 0.331 | 0.197 | 0.247 | |
| 2 | O+P | 0.408 | 0.317 | 0.357 | 0.110 |
| 3 | O+P+HN | 0.584 | 0.549 | 0.566 | 0.209 |
| 4 | O+P+HN+UE | 0.611 | 0.571 | 0.590 | 0.024 |
| 5 | O+P+UE+HN+A | 0.625 | 0.589 | 0.606 | 0.016 |
| 6 | O+P+UE+HN+BP | **0.626** | **0.596** | **0.611** | 0.020 |
| 7 | O+P+UE+HN+BP+A | 0.616 | 0.585 | 0.600 | -0.011 |
| 8 | O+P+UE+HN+ALLBP | 0.625 | 0.588 | 0.606 | -0.005 |

Table 6. The contribution of features and the progressive effects from adding more features.

moderately informative, only 4 NE categories are recognized in any way. NEs among the minor categories cannot be identified and most of the entities are classified as Protein, as most have the same surface appearance as protein names. The overall F-score achieved is 0.247. Addition of the POS features (P) provides limited information on the classifications. It leads to an increase of 0.110 on the F-score. The Head noun feature (HN) is very useful and provides a positive effect of 0.209 on the F-score compared to using simple Orthography plus POS tags. Adding in the unigram entity feature (UE) also provides a further improvement of 0.024 in F-value. These four features (O+P+HN+UE) are the most informative features, so if we treat these four features (Exp 4) as a new baseline for the remainder of the experiments we can discuss each other experiment relative to this baseline.

The Affix features (A) lead to a small positive improvement by 0.016 (Exp 5). Adding the bi-gram phrase feature (BP) (Exp 6) gives a slight increase in F-value (0.020). However combining affix features and bi-gram phrase features together (Exp 7) slightly degrades the performance by 0.011 and 0.006 respectively compared with Exp 6 and Exp 5. It may be that the affix and bi-gram phrase features carry some overlapping information, and contribute to some conflict. We assumed that more bi-gram phrase features can make more contribution to the classification, and performed experiments including low frequency bi-gram phrases (Exp 8), and the results shows some noise was introduced into classification with a slight F-value drop by 0.005.

## 6.2 Effect of post processing

The results of post-processing are reported in Table 7. Using the best model in the ME classification (Exp 6) as the baseline we applied tag changes and boundary correction to it. By using method 1 to fix the invalid I tags, the precision is degraded by 0.055, but with a moderate increase in recall (0.010) and decrease in F-score by 0.022 (Exp 10). Error analysis shows many false positives such as single lower case words. After correcting for invalid tag sequences (Exp 11) there is a slight increase in F-score (0.012). Next we used the Experiment 3 results as the second baseline on which to apply the boundary corrections.

The right boundary correction (Exp 12) further increases in F-score by 0.012 and the left boundary correction (Exp 13) also has a positive effect of 0.004. Applying both left and right boundary corrections there is a total increase of 0.015 in F-score. The left boundary correction only gives a slight positive effect, suggesting that the left boundary is more difficult to detect than the right boundary. The results of Experiment 15 show that the reclassification according to the head nouns has a positive effect on performance, improving the overall F-value by 0.044 compared to experiment 14.

The combined effect of post-processing is very effective, improving the performance over the ME model baseline by 0.071.

## 6.3 The Results

Table 8 shows the precision, recall and F-scores of the most populous categories of NE. The Protein class has the highest values for precision and recall. This is possibly due to proteins being the most frequent entity category in the training set. The Othername class is the second most frequent category, but it does not have a comparably high F-score. This is possibly due to the fact that Othername consists of some nested named entities which cause overlapping between Othername and other categories. Some small categories have very low F-score, due to a lack of training data.

| Category | P | R | F | Freq |
|---|---|---|---|---|
| **Protein** | 0.739 | 0.743 | 0.741 | 33.07% |
| **OtherName** | 0.653 | 0.643 | 0.648 | 25.43% |
| **DNA** | 0.718 | 0.642 | 0.678 | 11.69% |
| **CellType** | 0.758 | 0.714 | 0.735 | 8.09% |
| **CellLine** | 0.696 | 0.640 | 0.667 | 5.06% |
| **Lipid** | 0.654 | 0.464 | 0.543 | 2.26% |
| **Overall** | 0.700 | 0.666 | 0.682 | 100% |

Table 8. Performance of major entity categories.

| Exp. # | Processing | P | R | F | Effect | Baseline |
|---|---|---|---|---|---|---|
| 9 | **Baseline** | 0.626 | 0.596 | 0.611 | - | - |
| 10 | **Change invalid I to B** | 0.571 | 0.606 | 0.588 | -0.022 | 9 |
| 11 | **Fix invalid tag sequence** | 0.639 | 0.608 | 0.623 | 0.012 | 9 |
| 12 | **Right Boundary Correction** | 0.651 | 0.619 | 0.635 | 0.012 | 11 |
| 13 | **Left Boundary Correction** | 0.643 | 0.612 | 0.627 | 0.004 | 11 |
| 14 | **Boundary Correction on both Side** | 0.655 | 0.623 | 0.638 | 0.015 | 11 |
| 15 | **Fix tag sequence according to head nouns** | **0.700** | **0.666** | **0.682** | **0.044** | **14** |

Table 7. Effect of post-processing

The partial matching performance and identification performance are presented in Table 9 with the performance of exact match, left boundary correct, right boundary correct and identification only.

| Boundary Performance | P | R | F |
|---|---|---|---|
| Exact match | 0.700 | 0.666 | 0.682 |
| Left Boundary | 0.722 | 0.687 | 0.704 |
| Right Boundary | 0.739 | 0.703 | 0.721 |
| Identification Only | 0.802 | 0.762 | 0.781 |

Table 9. Partial matching and identification

The left boundary and right boundary have a higher performance than exact match, by .022 and .039 respectively in F-score. The results also show that the right boundary identification is better than left boundary identification. This shows that the left boundary is more difficult to detect, probably because of the difficulty in determining whether a modifier should be included in an NE or not. Identification outperforms classification by 0.099 F-value.

| | P | R | F |
|---|---|---|---|
| Shen et al.[9] | 0.677 | 0.653 | 0.665 |
| Lee et al.[6] | 0.718 | 0.698 | 0.708 |
| Zhou et al.[17] | 0.727 | 0.698 | 0.712 |
| Lin et al.[8] | 0.727 | 0.715 | 0.721 |
| Experimental System | 0.700 | 0.666 | 0.682 |

Table 10. A comparison to other systems

In Table 10 we show a comparison of our results to other systems. Although the test data is not exactly the same in each system, but for a rough comparison, our system achieved a performance close to these systems, and our system outperformed Shen's system slightly. Our system reported a relatively high boundary identification results, we think this is because the unigram entity feature and bi-gram phrase feature contributed to improve boundary identification.

## 6.4 Error analysis

Large numbers of misclassifications arise between the DNA and Protein classes. In the total of misclassified words, about 75% of the incorrectly recognized DNA terms are classified as Protein. This is due to the high overlap between these two classes. Another two categories that cause confusion is the CellLine of which 72% are incorrectly classified as CellType. These kinds of problems will most probably be addressed by exploring more contextual information.

Recognition error arises in some hyphen suffixes. For example, the entity "AP-1 –binding activity" of Othername has been partially recognized as AP-1 $_{\text{Protein}}$ –binding $_{\text{Outside}}$ activity $_{\text{Othername}}$. Similar situations are confronted with some high frequency hyphen suffixes, such as the word "cell-specific". This problem may be solved by a more careful study of hyphenated word features.

Abbreviation is another source of misclassification. The orthographic feature cannot capture enough information on abbreviations, because most abbreviations share the same orthographic feature. For example, the name "LPL" of Protein has always been recognized as Lipid.

True negatives are almost always identified by the feature of LowcaseOnly but are confounded with some entities. For example, the phrase "protein products" is never correctly labelled by our recognizer. These errors might be detected by using a dictionary, or exploration of more context information.

Other sources of errors are a number of non-entity words that are common medical terms classified as entities. Some high frequency words, such as stop words are incorrectly classified as Othername, as they sometimes appear in the composition of long entity names, for example, the word "family" and the word "and" have often been recognized as NE.

## 7 Conclusion and future work

In this paper we have presented a machine learning system for recognizing entity classes in biomedical abstracts. We have studied various linguistic features such as orthography, part of speech, affixes, head nouns, unigram terms and bigram phrases. We have also used simple rule based methods to correct invalid tag sequences and entity boundary errors.

We have achieved close to state of the art performance using very simple rule based post-processing without exploiting dictionaries. Our system achieved relatively high performance on boundary detection. However, there is still a 10% F-score gap between the identification performance and classification performance. This suggests that we have the potential to achieve better performance by looking at more informative features for semantic classification. In future work we will pursue better definitions of phrase forming rules and separate out the predictive value of different features for different entity types which is clearly shown to be operating in the use of the orthographic feature.

## References


[1]. N. Collier, C. Nobata, and J. Tsujii. *Extracting the names of genes and gene products with a hidden Markov model*. In Proceedings of *COLING 2000*, pp 201-207, 2000.

[2]. N. Collier, K. Takeuchi. *Comparison of character-level and part of speech features for name recognition in biomedical texts.* J Biom. Inform. 37. pp423-435. 2004.

[3]. K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. 1998. *Toward information extraction: identifying protein names from biological papers. In Proc. of the Pacific Symposium on Biocomputing'98 (PSB'98)*, pp 707-718..

[4]. R. Gaizauskas, G. Demetriou and K. Humphreys. *Term Recognition and Classification in Biological Science Journal Articles. 2000*. In Proc. of *the Computional Terminology for Medical and Biological Applications Workshop* pp 37-44. 2000

[5]. L. Hirschman, A.A. Morgan, and A.S. Yeh, *Rutabaga by any other name: extracting biological names. J Biomed Inform*, 2002. 35(4): p. 247-59.

[6]. J. Kazama, T. Makino, Y. Ohta, J. Tsujii. *Tuning Support Vector Machines for Biomedical Named Entity Recognition.* In: Proceedings of *Workshop on NLP in the Biomedical Domain, ACL 2002*. pp1-8. 2002.

[7]. K.-J. Lee, Y.-S. Hwang, and H.-C. Rim. *Two-phase biomedical NER recognition based on SVMs*. In *Proceedings of ACL 2003, 2003.*

[8]. Y. Lin, T. Tsai, W. Chou, K. Wu, T. Sung and W. Hsu: *A Maximum Entropy Approach to Biomedical Named Entity Recognition*, In: Proceeding of *the 4th Workshop on Data Mining in Bioinformatics*: pp 56-61, 2004

[9]. D. Shen, J. Zhang, G. Zhou, S. Jian and L. Tan, *Effective Adaptation of a Hidden Markov Model-based Named Entity Recognizer for Biomedical Domain*, In: Proceedings of ACL *2003 Workshop on NLP in Biomedicine, Sapporo, Japan*, pp49-56, 2003.

[10]. B. Settles. *Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets*. In: Proceedings of *the COLING 2004 NLPBA,*. 2004, pp 104-108, 2004.

[11]. B. Sundheim *Overview of the results of the MUC-6 evaluation*. In: Proceedings of *the Sixth Message Understanding Conference. Los Altos, CA: Morgan Kaufman; 1995*. p. 13-31

[12]. K. Takeuchi, and N. Collier. *Bio-medical Entity Extraction using Support Vector Machines.* In: Proceedings of *NLP in Biomedicine, ACL 2003. Sapporo, Japan*, pp 57-64, 2003.

[13]. Tsai, T.-H., Wu, S.-H., & Hsu, W.-L. (2005). *Exploitation of linguistic features using a CRF-based biomedical named entity recognizer.* to appear in ACL Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics, Detroit

[14]. Y. Tsuruoka, Y. Tateishi, . Kim, T. Ohta, J. McNaught, S. Ananiadou, and J. Tsujii, *Developing a Robust Part-of-Speech Tagger for Biomedical Text*, Proceedings of *the 10th Panhellenic Conference on Informatics, 2005*

[15]. Tuason, O., L. Chen, H. Liu, J.A. Blake, and C. Friedman. *Biological Nomenclature: A Source of Lexical Knowledge and Ambiguity*. In: Proceedings of *Pac Symp Biocomput*. 2004. p. 238-49.

[16]. G. Zhou, *Recognizing Names in Biomedical Texts using Hidden Markov Model and SVM plus Sigmoid*, In: In: Proceedings of *the COLING 2004 NLPBA, Geneva, Switzerland*. 2004.

[17]. G. Zhou and J. Su. *Named Entity Recognition using an HMM-based Chunk Tagger*. In Proc. of *the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 473-48*0 2002.

[18]. C. Nobata, N. Collier and J. Tsujii. *Automatic term identification and classification in biology texts*. In Proc. of *the 5th NLPRS, pp 369-374*. 1999.

[19]. M. Krauthammer, Rzhetsky A, Morozov P, Friedman C. *Using BLAST for identifying gene and protein names* in journal articles. Gene 2000;259(1–2):245–52. 2000

[20]. J. Kim, T. Ohta, Y. Teteisi, and J. Tsujii. *GENIA corpus – a semantically annotated corpus for bio-textmining. Bioinformatics 19 (suppl.1)* 2003

# Evaluating an ontology with OntoClean

*Jonathan Yu* *James A. Thom* *Audrey Tam*

School of Computer Science and I.T., RMIT
124 La Trobe Street, Melbourne, Victoria 3000, Australia

Email: {jyu,jat,amt}@cs.rmit.edu.au

**Aim:** To apply the OntoClean methodology to a subset of an ontology in the travel domain and assess the strengths and weaknesses of the methodology based on the resulting ontology.

**Methods:** The modelled concepts and relations in an ontology provides a common understanding of a domain for parties to agree or commit to. Thus ontologies can be used to facilitate interoperability between applications.

This OntoClean methodology can be used to verify an ontology for its correctness. Correctness refers to whether the modelled entities and properties in an ontology correctly represents entities in the world being modelled. This methodology examines the usage of the subsumption relation between classes in an ontology using formal notions (rigidity, unity, identity and dependance) to capture various characteristics of classes, and constraints upon those metaproperties [1]. We applied this to a subset of the Lonely Planet (LP)[1] ontology describing activities. LP uses the ontology to organise the structure and contents of documents for travel guides.

**Results:** We can observe an improvement from the original ontology after applying OntoClean outlined in the resulting ontology as shown in Figure 1 and 2 respectively. However, a limitation we found with this methodology is that the resulting ontology may not have the best structure. That is, it does not guide as to how it can be further structured. For example, the tour activity branch of the ontology can be further grouped into *Guide-oriented*, *Interest-oriented* and *Transport-oriented* categories.

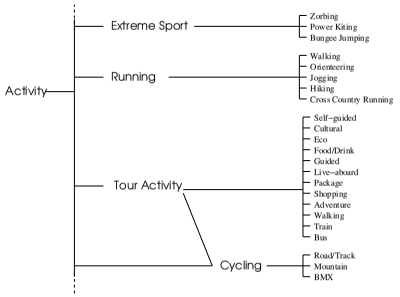



*Figure 1.* Sections of original LP activity ontology

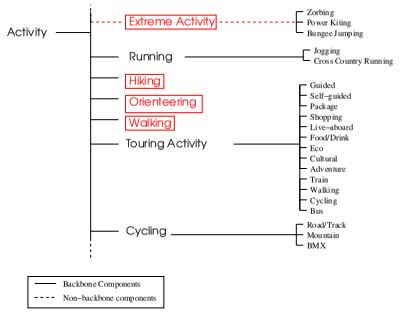



*Figure 2.* Sections of resulting LP activity ontology after OntoClean

**Conclusions:** We have considered the OntoClean methodology in this study and applied it to a subset of an ontology in the travel domain – the Lonely Planet activity ontology. Using the OntoClean metaproperties, it was found that modelling assumptions were clarified and inconsistencies were discovered. Thus we were able to correct these inconsistencies. However, it was noted that the resulting ontology could be further improved.

## References

[1] N. Guarino and C. Welty. Evaluating ontological decisions with ontoclean. *Communications of ACM*, Volume 45, Number 2, pages 61–65, 2002.

[1] http://www.lonelyplanet.com

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.**

# Document Ranking for Effectiveness-Efficiency Tradeoffs

*Vo Ngoc Anh* *Alistair Moffat*

Department of Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

{*vo,alistair*}*@csse.unimelb.edu.au*

**Aim:** A large-scale document ranking system should be both effective and efficient. Here we summarize the main features of a prototype system we built for that purpose. The success of the system is demonstrated through its relative performance for the efficiency task of the TREC 2005 Terabyte Track.

**Method:** Building a document ranking system involves two key decisions: choosing a retrieval model, and choosing a suitable index representation. The former determines the *effectiveness* of the system, the latter the *efficiency*; and each of them affects the other.

The impact-based document ranking mechanism described by Anh and Moffat [2] was chosen for our system because of its balance between effectiveness and efficiency. In terms of effectiveness, it is highly competitive, although still inferior to advanced language modelling implementations. On the other hand, in terms of efficiency the mechanism is excellent, as it ranks documents using a small number of calculations, all on integer numbers. To further facilitate query efficiency, we compress the index using the `slide-8` coding scheme [1], which allows an excellent balance between compression ratio and decoding speed.

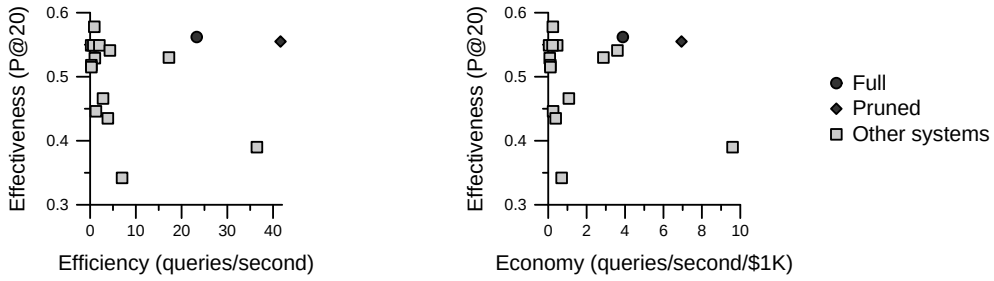



Figure 1: Relative performance in the 2005 TREC Terabyte Track (using the 426 GB `GOV2` collection), with effectiveness compared to efficiency (left-hand graph) and economy (right-hand graph), the latter calculated as throughput normalized by estimated system cost ($US). Efficiency is measured over 50,000 real-life queries; effectiveness over a subset of 50 queries.

**Results:** Figure 1 shows the relative performance of our system in the efficiency task of the TREC 2005 Terabyte Track. The graphs were compiled from data covering the 16 runs submitted by the 8 groups with the best scores according to the metric P@20 [3]. Our two runs are labelled `Full` and `Pruned`. The former refers to the full processing of all pointers in all inverted lists, the latter to a run in which low-impact pointers were ignored.

**Conclusion:** This year's involvement in the TREC Terabyte Track showed that our system provides a good balance between effectiveness and efficiency.

## References

[1] V. N. Anh and A. Moffat. Improved word-aligned binary compression for text indexing. Submitted, 2005.

[2] V. N. Anh and A. Moffat. Simplified similarity scoring using term ranks. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates and N. Ziviani (editors), *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 226–233, Salvador, Brazil, August 2005. ACM Press, New York.

[3] C. L. A. Clarke and F. Scholer. The TREC 2005 Terabyte Track. In *The Fourthteenth Text REtrieval Conference (TREC 2005) Notebook*, Gaithersburg, MD, November 2005. National Institute of Standards and Technology. Available at http://trec.nist.gov/act_part/t14_notebook/t14.notebook.html.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

# Document Priors for Query Prediction

Steven Garcia  Nicholas Lester  Justin Zobel
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne 3001, Australia
{garcias,nml,jz}@cs.rmit.edu.au

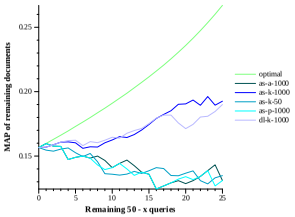
**Aim:** To predict the difficulty of a query issued to an information retrieval system. A query is considered difficult when the effectiveness of the search results are poor.

**Methods:** It has been shown that the likelihood of document access from a collection is non-uniform [2]. As such, for each document in a collection, a probability can be obtained that indicates the likelihood of seeing that document in any given result set. We propose several approaches to query difficulty prediction that take advantage of the non-uniform likelihood of document access by a search system.

Given a set of document probabilities, an absolute ordering of documents from most to least likely to be retrieved indicates a default ranking for documents in the collection. For a generic query, we expect the documents in the result set to be ranked in much the same order as the absolute ordering based on the document prior probabilities. Therefore, a query that produces a result set with documents that do not significantly differ in order to the prior based absolute ordering, is considered to be a difficult query. Conversely, a query that produces a result set that significantly differs in order to the absolute ordering is considered to have high discriminatory power, and therefore is considered a simple query to resolve. We propose five query prediction measures based on document priors, discussion of each technique is presented elsewhere [1].

In recent years, query difficulty prediction has been incorporated into TREC as a part of the Robust track [3]. One metric to measure the quality of a set of predictions is the area under the curve metric that measures the quality of the prediction by measuring the difference in average precision between the worst 25 predicted topics of a run, to the worst 25 performing topics. We use this metric on the TREC 2005 Robust topics and the Aquaint collection.

**Results:** The figure illustrates the performance of our five techniques using the area under the curve approach. Each line shows the mean average precision for the remaining TREC queries after removing the worst $x$ queries. The best possible prediction for this run is shown with the optimal curve. The area between the optimal curve and the other curves shows the gap between that approach and the optimal prediction.




**Conclusions:** We explore a novel approach to query difficulty prediction and propose five metrics to determine the query difficulty based on document priors. Two of the five techniques show promise as predictors. We plan to further explore difficulty prediction using document priors in combination with other query prediction techniques, with the hope of further improving query difficulty prediction.

## References

[1] Y. Bernstein, B. Billerbeck, S. Garcia, N. Lester, F. Scholer, J. Zobel and W. Webber. Rmit university at trec 2005: Terabyte and robust track. In E. M. Voorhees and L. P. Buckland (editors), *Proc. Text Retrieval Conf. (TREC)*, Gaithersburg, MD, November 2005. National Institute of Standards and Technology. Proceedings to appear.

[2] S. Garcia, H. E. Williams and A. Cannane. Access-ordered indexes. In V. Estivill-Castro (editor), *Proceedings of the 27th Conference on Australasian Computer Science*, Volume 26, pages 7–14, Dunedin, New Zealand, January 2004. Australian Computer Society.

[3] E. M. Voorhees. Overview of the TREC 2004 robust track. In E. M. Voorhees and L. P. Buckland (editors), *Proc. Text Retrieval Conf. (TREC)*, Gaithersburg, MD, November 2004. National Institute of Standards and Technology Special Publication 500-261.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

# Information Retrieval Estimation via Fuzzy Probability

*Zhiheng Huang*
Department of Computer Science
The Australian National University
Australia
*zhiheng@cs.anu.edu.au*

*Tamás D. Gedeon*
Department of Computer Science
The Australian National University
Australia
*tom@cs.anu.edu.au*

**Aim:** Fuzzy logic [3] is a useful approach to help identify partially matched documents for a given query. Research on fuzzy information retrieval mainly focuses on the relevance probability estimation of the retrieved documents but lacks the estimation of imprecision of such probability. This study attempts to use fuzzy probability to evaluate the relevance of retrieved documents.

**Methods:** Probability is widely used in information retrieval to estimate the relevance of retrieved documents. As *relevance* cannot be defined accurately, i.e., it is subjective in practice, the probability estimate of the retrieved documents cannot be precise. This suggests the use of an imprecise representation of probability estimation, which is termed *fuzzy probability*.

One example (as Fig. 1) is to show the fuzzy probability of a document to be "very relevant" to a query. As the definition of "very relevant" is imprecise, the probability of such a retrieved document is hence imprecise. As can be seen, it is *certain* ($possibility = 1$) for such a document to be "very relevant" with a probability of $30\%$, and it is *quite certain* ($possibility = 0.8$) for it to be "very relevant" with a range of probability from $20\%$ to $35\%$. The higher possibility value leads to the narrower estimation of probability. In the extreme case, the peak point in the fuzzy probability corresponds to the conventional probability estimate.

There are considerable studies on fuzzy probability. For example, the possibility-probability distribution [1] has been recently proposed. The problem of decision making [2] in the face of a fuzzy probability estimate is investigated.

**Results:** We have proposed a novel method to calculate fuzzy probability. This method will be applied to calculate a document's relevance (in terms of fuzzy probability) for a query. By using the calculated fuzzy probabilities, different crisp probabilities can be computed [2] with respect to different uncertainty levels (such as *certain* and *quite certain*), which provides a flexible way to evaluate the relevance of a given document.

**Conclusions:** This study has proposed a novel method to calculate fuzzy probability. At this stage, no experimental work has yet been carried out. The experimental evaluation of this method in information retrieval will be our most important future work.

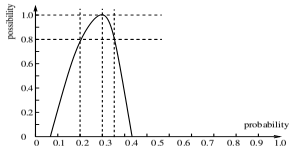



Figure 1: A fuzzy probability estimation of a document to be "very relevant"

## References

[1] C. F. Huang and Y. Xue. Some concepts and methods of information granule diffusion. In *IEEE International Conference on Granular Computing*, Volume 1, pages 28 – 33, 2005.

[2] R. R. Yager. Decision making with fuzzy probability assessments. *IEEE Transactions on Fuzzy Systems*, Volume 7, 1999.

[3] L. A. Zadeh. Fuzzy sets. *Information and Control*, Volume 8, 1965.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.**

# A Metadata Collection Technique for Documents in WinFS

*Stijn Dekeyser*

Department of Maths & Computing, University of Southern Queensland (USQ), Australia

**Aim:** To propose a relatively preliminary yet sufficiently general GUI-based technique to collect rich metadata for documents, to be used in a file system wrapper such as Microsoft's forthcoming WinFS.

**Methods:** With the recent beta release of Microsoft's WinFS file system wrapper, the traditionally disjunct research areas of databases, document computing, and file systems are merging. In such environments, it is relatively clear how to query the file system for metadata about documents. However, it is not yet clear how rich metadata can be collected for documents in an intuitive manner suitable for novice users. We present a GUI-based technique for capturing both *formative* and *contextual* metadata. The former relates to metadata that can be set automatically when a document is created (such as the shutter-speed setting for a digital camera image), while the latter represents the much richer metadata that can only be determined through user interaction (e.g. the persons appearing in the image). The technique, described in more detail in [2] entails three parts:

• **Class Hierarchy.** We propose to employ a hierarchical inheritance-based class diagram that complements WinFS's built-in technically-oriented classes. The hierarchy is built in three phases: first, we propose that the first few levels in the hierarchy contain built-in classes that represent tangible concepts (e.g. Person, Product etc.) likely to be of general use. Secondly, we propose that businesses create classes in the next few levels of the hierarchy representing concepts in the company's specific branch of work (e.g. Insurance), and make those classes available in staff's computers. Finally, the lowest levels of the hierarchy should contain classes created by individual users (e.g. Flooding) to increase their productivity. For all newly created classes, their *properties* should reflect formative metadata while their *relationships* to other classes should reflect contextual metadata. We further propose straightforward, intuitive GUI actions to create new classes and their properties and relationships.

• **Capturing formative metadata.** A newly created class will be represented by an icon that acts as a folder for objects of that class. Creating a new document thus is simply adding an object to the folder; formative metadata is automatically provided by the class' constructor method (e.g. setting an image's metadata can be done through reading its EXIF information when copying it from the camera).

• **Capturing contextual metadata.** In WinFS all users can create relational views over documents and their metadata. We propose to create Class Folders which are updatable views [1] usually over just one table (class), although more complex ones are possible (see [2]). Such views are homogeneous (although taking inheritance into account) and in effect become virtual classes, in which objects and their formative metadata can be created as before. As our proposal ties contextual metadata to relationships between classes, our approach to collecting it upon saving is to supply the user with an Explorer-like window that contains several tabs. Because the document belongs to a certain class, we know which relationships to other classes are associated to it. Thus, we create one tab in the save-as dialog window per relationship. Inside those tabs, all member documents of the target class are displayed, together with all Folder views defined on that target class. The user is then asked to, *per tab* (i.e., per relationship of which the document's class is a source), select the documents and/or Folders to which the document is associated through that specific relationship. This selection procedure is akin to the current "Save As" dialog (although may take longer for the user to perform depending on the number of relationships) and serves to visually capture the contextual metadata rather than requiring its input through keyboard entry.

**Results:** This paper presents the concepts involved in our metadata collection technique rather than a working prototype. Hence we do not yet have but are planning to obtain results measuring usability (e.g. through ISO 9241 metrics).

**Conclusions:** We have proposed a generic technique to capture rich metadata for documents in the context of WinFS. The technique is based on sound relational theory and extends current "File Save As" GUI dialogs to assist users in capturing complex metadata for user-defined types in a reasonable, intuitive manner. We plan to implement a proof-of-concept system to test usability and effectiveness.

## References

[1] U. Dayal and P. Bernstein. The updatability of relational views. In *VLDB'78*, pages 368–377, 1978.

[2] S. Dekeyser. Metadata collection for documents in WinFS. Technical Report SC-MC-0524, USQ.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

# Hosting search services for the Australian Government

*George Ferizis*
CSIRO ICT Centre
ACT 2601 Australia
*George.Ferizis@csiro.au*

*David Hawking*
CSIRO ICT Centre
ACT 2601 Australia
*David.Hawking@csiro.au*

**Aim:** CSIRO's information retrieval group currently host several search related services for the Australian government (accessible, for example, through *http://www.australia.gov.au*), using the P@noptic search engine (*http://www.panopticsearch.com*). We aim to address issues such as the reliability and cost effectiveness of the services and the quality of the search results.

**Methods:** The reliability of the service has been increased by the introduction of fault tolerance. Several servers serve queries for each service so that if one server is offline the other servers will continue serving queries.

We make measurements on the quality of the search results using various evidence in the document, including *metadata*, *document title*, *anchor text*, *document contents* and *click data*, to determine what evidence returns the most relevant results. We also measure the benefits of an additional, smaller daily crawl to supplement a weekly large whole of government crawl. These are measured by obtaining the mean rank of the homepage of government agencies when queries containing the name of the agency are submitted.

We also assess the bandwidth costs associated with a large whole of government crawl. To reduce bandwidth costs we use a technique named *"incremental crawling"*, which compares the content length present in the HTTP header returned by a web server and the length of the document from previous crawls. If the length has not changed the document is not downloaded again. We measure the reduction in the data downloaded using this method.

**Results:** Figure 1, shows the effects of using different evidence for ranking. The mean reciprocal rank(MRR) that is shown is defined as the average reciprocal rank of the "correct" result to a query over several different queries. The results show that using evidence provided by a reader of the document (eg. anchor text, or search user clicks) gives more relevant results thatn using any evidence the document provides.

The network traffic reduction that results from the use of incremental crawling show that it is possible to obtain a content length header for approximately 30% of the web pages crawled, and approximately 90% of these have not changed from the previous crawl. On a recent crawl of federal government web sites it was found that incremental crawling reduced the amount of data downloaded from 140 gigabytes to 60 gigabytes.

**Conclusions:** Improving accessibility of government services and information to the public has provided a great opportunity to deliver impact from our research. It has also posed an interesting and diverse set of new engineering and scientific challenges. By studying and addressing customer requirements in the areas of search quality, functionality, coverage, freshness, efficiency, robustness, and cost-effectiveness, we have improved both our technology and our understanding.

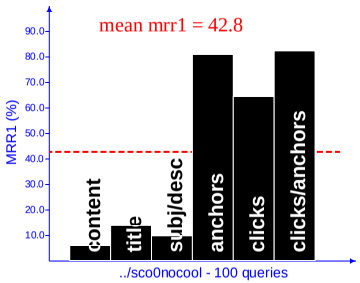



Figure 1: Effects of different evidence on ranking quality

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 

# Automatic Identification of English and Indonesian Parallel Documents

Jelita Asian Falk Scholer S.M.M. Tahaghoghi Justin Zobel
Email: {jelita,fscholer,saied,jz}@cs.rmit.edu.au
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne 3001, Australia.

**Aim:** Parallel corpora are useful for cross-lingual information retrieval and other natural language processing tasks. However, current techniques for finding parallel documents rely on file names and file structures, and semantic and statistical information in documents. We aim to automate the identification of English and Indonesian parallel documents.

**Methods:** We have developed a new global alignment method, based on the methods used for applications such as matching of protein sequences, to align windows of words between documents and queries. The documents can be translated to the query language before the alignment, or – as is the case for Indonesian and English – they can remain untranslated if they share the same character set. For each pair of documents, we group words into windows of a certain size with a 50% overlap with the next window. We then count the number of unique words in common between the windows. We use a global alignment method for aligning protein sequences, based on the Needleman and Wunsch algorithm [1], to align these windows of words. The basic principle of the method is to find as many matches as possible between the two documents, and to punish when there is any insertion or deletion.

**Results:** We compare our alignment schemes, using different window sizes and penalty values, with results obtained by documents indexed using a search engine called Zettair. The following table shows that our alignment method can differentiate between parallel and non-parallel documents when compared to a search engine baseline. This differentiation is measured using a separation value, which is the difference between a highest false match and a lowest true match. The negative separation value indicates that the Zettair baseline ranks non-parallel documents higher than parallel documents. A smaller window size works well for untranslated English documents, while a larger window size works well for translated English documents.

| | | Untranslated | Translated |
|---|---|---|---|
| Zettair Baseline | | -24.7097 | -0.2471 |
| | Window Size | | |
| Optimum scheme | 12 | 27.7194 | — |
| | 28 | — | 19.1492 |

**Conclusions:** Our global alignment method is successful in separating parallel documents between Indonesian documents with either translated or untranslated English documents.

## References

[1] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, Volume 48, Number 3, pages 443–453, 1970.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.**

# SciFly - Customised Flyers on Demand

*Andrew Lampert*

Information Engineering Laboratory
CSIRO ICT Centre
North Ryde NSW Australia

*Andrew.Lampert@csiro.au*

**Aim:** To demonstrate how the delivery of information can be improved beyond lists of search results by reasoning about the context of a user's interaction and adapting both the content and presentation accordingly.

**Methods:** SciFly is a demonstration system that generates customised flyers about CSIRO's research in Information and Communication Technologies, based on user-selected areas of interest. It has been designed to operate as a touch-screen information kiosk, but also has a web-interface to allow remote, browser-based interaction.

In response to a user nominating their areas of interest, SciFly dynamically gathers and assembles relevant content into a flyer. SciFly also adds relevant contact information, web links and higher level context, all of which is tailored to support the information presented. This process is controlled using the flexible planning capabilities of our Myriad delivery platform (see Paris et al. [2] for details).

The automatically generated flyers are modelled on existing manually authored flyers. These manual flyers are professionally authored and pre-printed in bulk, and as such are static and expensive to update. To inform the content and presentation of our automatically tailored flyers, a corpus analysis of human-authored flyers was performed to understand both their structure and content. This knowledge is encoded in the rules that configure the Myriad planning engine, which controls both the retrieval and delivery of information.

The assembled information for each flyer is structured according to the rhetorical relations that exist between segments of text, as described by Mann & Thompson [1]. Importantly, SciFly does not just present pre-configured content, but dynamically adjusts the amount and detail of content based on the range of topics selected, the structure of the information to be presented and the constraints of the delivery medium. In particular, the rhetorical structure information allows SciFly to ensure that the information presented remains coherent and consistent even when adapted to the available space and different output devices.

**Results:** SciFly simultaneously prints and emails the tailored flyer to the user, along with a condensed, plain text summary of the flyer in the body of the email message. Thus users receive information relevant to their interests, the presentation of which is adapted to suit the different output devices. Specifically, users can receive a:

- **Double-sided paper flyer** as a physical flyer for immediate perusal;
- **PDF document** for electronic delivery and later reference; and
- **Plain text summary** for mobile device access.

We are in the process of evaluating the value of both the coherence and content tailoring within SciFly.

**Conclusions:** The algorithms and ideas embodied in the SciFly application demonstrate how the delivery of information can be improved beyond the ranked lists of results that are common in many search engines when more is known about the context of a user's interaction. In this case the context reasoned about includes user preferences, likely task and the delivery device.

## References

[1] William Mann and Sandra Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, Volume 8, pages 243–280, 1988.

[2] Cecile Paris, Keith Vander Linden, Matt Post and Shijian Lu. Myriad: An architecture for contextualized information retrieval and delivery. In *AH2004: International Conference on Adaptive Hypermedia and Adaptive Web-based Systems*, pages 205–214, 2004.

**Proceedings of the 10th Australasian Document Computing Symposium, Sydney, Australia, December 12, 2005.** 