# TREC 2007 Enterprise track at CSIRO

*Peter Bailey*
CSIRO ICT Centre
Canberra, Australia
*peter.bailey@csiro.au*

*Deepak Agrawal*
CSIRO ICT Centre
Canberra, Australia
*deepak.cse87@gmail.com*

*Anuj Kumar*
CSIRO ICT Centre
Canberra, Australia
*anuj_kumar@iiitm.ac.in*

October 16, 2007

## 1 Introduction

The goals of CSIRO's participation in the Enterprise track were formed by the nature of the tasks. With the expert finding search task, we sought to use a variety of means to associate topical expertise with individuals previously located within the collection. With the document search task, we were primarily interested in exploring issues of result diversity based on different characterisations of documents within the collection.

We completed both expert and document search tasks by the submission deadline. In both cases, we submitted four runs for each task. The algorithms used for the runs for both tasks used a query-only baseline with subsequent variations. In both cases, we incorporated use of the PADRE retrieval system [2], in which the Okapi BM25 relevance function was implemented as the core ranking component. Incorporation of additional evidence such as anchor text and other characteristics of Web documents is used in the default ranking formula associated with the retrieval system.

## 2 Expert search task

The expert search task was substantially different to prior years in the Enterprise track. First, judgments of (a small list of) who the key contacts were for a topic were made by the science communicators who created the topics in the first instance. Science communicators are people employeed within CSIRO to communicate science to the public. This meant that the task was more oriented towards a system with high early precision than recall. Second, there was no list of candidates provided in advance. Candidate experts were to be identified (somehow) within the corpus, and then reported by email address, as a form of unique id.

### 2.1 Expert search pre-processing

The first challenge was thus to process the document collection and attempt to discover email addresses from it. Normal CSIRO email addresses for individuals are in the form *Firstname.Lastname@csiro.au*. However, case formatting of email addresses is inconsistent. And in many cases, email addresses are obfuscated through the use of character encodings for the `@` symbol. These alternate forms of email addresses had to be identified, extracted and canonicalised. Subsequently, a reversible mapping of canonical email uids to documents within the collection was made. From the email uids, it was straightforward to generate a set of people's names.

The corpus was analysed to determine that certain URL patterns corresponded to home pages. The vast majority of these were found as profile pages on *www.csiro.au/people/* and *www.ict.csiro.au/staff/*. A notional sub-corpus was created from these documents — *profiles*. A standard PADRE index was created over this sub-corpus.

Another notional sub-corpus — *surrogates* — was created by querying PADRE over an index of the full corpus with the names of people identified earlier. For each person, a surrogate document corresponding to the PADRE HTML output format for the top 50 results for a query of the form "Firstname Lastname" was generated. e.g. An expert previously identified through the email uid as *Peter.Bailey@csiro.au* would become a phrase search query "Peter Bailey". It was hoped that the 5-line query-biased summaries generated for each result would contain relevant topical information in close proximity to occurrences of the person's name. Again, a standard PADRE index was created over this sub-corpus.

Two additional methods involving anchor text were involved. It is known that anchor text is a strong indicator of topical relevance from prior work (e.g. [6]).

The first — *expert anchor terms* — calculated *tf.idf* scores over the anchor texts associated with individual documents found in the full corpus. These documents were ranked according to the summation of these scores over the query terms on a per-topic basis. Then the experts associated with each document were scored correspondingly (for both the expert names and the expert email uids found in association), and summed. Concretely, if a document had anchor text associated with it of "information retrieval", then *tf.idf* scores for the terms "information" and "retrieval" were calculated, with respect to all anchor text available. Then if an expert name "Peter Bailey" was associated with the document, and the topic query was "information retrieval", the corresponding expert uid *Peter.Bailey@csiro.au* would receive the sum of the *tf.idf* scores associated with "information" and "retrieval". If the expert email address *Peter.Bailey@csiro.au* was also associated with the document, it too would get additional scores. Finally an expert list was reported in score order. Note that this method does not discriminate against partial matches of the query terms.

The second — *expert document match* — used PADRE to find exact matches of the query in the anchor text associated with documents. Concretely, if a query consisted of "information retrieval" then those terms must match exactly as a phrase in anchor text associated with a document for it to be included. Then the same process was carried out of mapping documents to experts (and their email uids), with the expert score being calculated on the basis of the frequency of email uids and names in all the documents returned, and the expert list being sorted in score order and reported.

## 2.2 Expert search runs

We submitted four runs for expert search, making use of different query input and combining our evidence in different ways. The runs, and a summary of their methods, are given in 1.

Table 2 shows the performance of each run. The baseline run is *CSIROesQonly*. Adding additional information available in the narrative run *CSIROesQnarr* improves performance marginally. The two other runs are particularly interesting.

First, in *CSIROesQprof* using only the pages available from the *profiles* dramatically degrades performance. (We believe that the use of the term 'expert' should not substantially affect the results due to PADRE's ranking mechanisms; documents which contain the word 'expert' would have been ranked higher, but if no documents contain the work 'expert' then results would be the same as if the term had not been included in the query to PADRE.) We speculate this poor result arises because insufficient information about "topicality" is available in the profile pages alone; other data in the rest of the corpus is important to help establish expertise.

Second, in *CSIROesQpage*, although we use only the *surrogates* sub-corpus, the additional information of key pages (analogous to a degree to click logs) results in the best performance of all 4 runs we submitted. The modification to the ranking algorithm (shown in Table 1) incorporates the key pages nominated by the science communicator for the topic as a way to up-weight the scores of experts for whom those key pages occur in their surrogate documents. This result confirms current research indicating the utility of external evidence for improved ranking performance [3], but at one level removed in the search task. Further investigation is merited here.

Table 1: Indexing and retrieval settings for the expert search task.

| Run ID | Indices | Query | Weighting |
|---|---|---|---|
| CSIROesQonly | surrogates, profiles, expert anchor terms, expert document match | Topic query only | A simple weighted combining algorithm was used, which summed the normalised scores. The profiles and expert document match scores were given higher weights. |
| CSIROesQnarr | surrogates, profiles, expert anchor terms, expert document match | Topic query as a phrase, each query term (mandatory), and the topic's narrative terms (optional) (with stopwords removed from both query and narrative) | As for CSIROesQonly |
| CSIROesQprof | profiles | As for CSIROesQnarr plus the addition of the term 'expert' | Normal PADRE weighting |
| CSIROesQpage | surrogates | Topic query only | If the topic key page URLs are found in a surrogate document, then $PADREscore \times (1 + \frac{1}{rank\ of\ URL\ in\ surrogate})$; otherwise $PADREscore$ |

Table 2: System performance for the expert search task.

| Run ID | MRR (gain) | MAP |
|---|---:|---:|
| CSIROesQonly | 0.5310 | 0.3517 |
| CSIROesQnarr | 0.5420 | 0.3655 |
| CSIROesQprof | 0.2564 | 0.1232 |
| CSIROesQpage | 0.5722 | 0.3660 |

Table 3: Indexing and retrieval settings for the document search task.

| Run ID | Query | Combining |
|---|---|---|
| CSIROdsQonly | Topic query only | Genre and PFCM results combined by Algorithm B |
| CSIROdsQnarr | Topic query and the topic's narrative terms (with stopwords removed from both query and narrative) | As for CSIROesQonly |
| CSIROdsQfb | Topic query | The key pages from the topic are used to adjust the weighting assigned to individual categories from the genre and clustering analysis; diversification is the goal |
| CSIROdsQsimp | Topic query | The key pages from the topic are used to adjust the weighting assigned to individual categories, such that pages which are similar to those already chosen are ranked higher |

# 3 Document search task

The document search task was a fairly standard Web search activity. Judging measures for the task are likely to reward successful early identification of key pages. The NDCG metric will be best suited for this task; however, results at the time of this paper are only available for 43 topics, and do not include NDCG.

## 3.1 Document search pre-processing

The system effect we were investigating was variations around the concept of result diversity [5, 1]. We speculated that CSIRO science communicators would prefer a variety of highly relevant documents (e.g. media releases, project pages, reports, profile pages, ...) not simply a straightforward ranked list.

We used two kinds of analysis to assist diversification. The first was some genre analysis, carried out in a topic-independent manner. The nature of the CSIRO corpus allowed us to carry out genre identification into a small number of interesting categories (people, projects, media releases, publications, biographies, feature articles, podcasts), using some simple regular expression matches over URLs and document texts. Documents were only allowed to appear in one category. Each category formed a separate pool of documents. Those documents which could not be categorised into any of these categories formed an additional pool of uncategorisable documents.

The second analysis we used was a proximity-based fuzzy clustering algorithm (PFCM) [4]. This was computed on a topic-dependent basis, over the pool of uncategorisable documents. Each clustering resulted in a new set of pools of documents. Documents were permitted to be in multiple clusters for a single topic.

## 3.2 Document search runs

We trialled three combination algorithms for selecting documents from a diversity of categories/clusters, and settled on one we called Algorithm B. This computed a score per document of $\sum_i \frac{non-zeromembershipofcategory(i)}{rank(i) \times globalRank}$ where *globalRank* is the rank of the document given by PADRE when it runs over the whole corpus.

Table 4: System performance for the document search task.

| Run ID | MRR (gain) | MAP |
|---|---:|---:|
| CSIROdsQonly | 0.8070 | 0.1800 |
| CSIROdsQnarr | 0.7258 | 0.1148 |
| CSIROdsQfb | 0.7962 | 0.1774 |
| CSIROdsQsimp | 0.7329 | 0.1660 |

The run *CSIROdsQonly* performed the best overall of any of the runs. The addition of query terms from the narrative in *CSIROdsQnarr* noticeably degraded performance for both MRR and MAP. Interestingly, for the runs making use of feedback information (the key pages provided by the science communicators in the topics), the diversity based approach of *CSIROdsQfb* performed better than selecting for similar pages in *CSIROdsQsimp*.

## 4 Discussion

For document search, the *CSIROdsQonly* run making use of only the topic's query information appears to be the most effective technique. The runs making use of the narrative information or the feedback (in the form of the key pages) appear to have hurt both MRR and MAP somewhat. Until the NDCG metric is calculated, it will be difficult to tell if our approach to diversification was successful. MRR results are encouraging, but the goal in this task is really to retrieve a few high value documents early in the list, not just one. The MAP measure does not bode well on this front. Comparison back to a native PADRE ranking would be worthwhile.

For expert search, the results are encouraging and deserve closer investigation. Additional work to find other sources of external evidence around topical associations to experts could bear fruit in this context.

## 5 Acknowledgments

We thank our colleagues (especially David Hawking) for invaluable advice throughout the development and experimental phases of this work.

## References

[1] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, New York, NY, USA, 1998. ACM Press.

[2] D. Hawking, N. Craswell, and P. Thistlewaite. ACYSys TREC-7 Experiments. In *Proceedings of Seventh Text Retrieval Conference*, Gaithersburg, USA, November 1998.

[3] D. Hawking, T. Rowlands, and M. Adcock. Improving rankings in small-scale web search using click-implied descriptions. *Australian Journal of Intelligent Information Processing Systems. ADCS 2006 special issue.*, 9(2):17–24, December 2006.

[4] V. Loia, W. Pedrycz, and S. Senatore. P-FCM: a proximity-based fuzzy clustering for user-centered web applications. *International Journal of Approximate Reasoning*, 34(2):121–144.

[5] F. Radlinski and S. Dumais. Improving personalized web search using result diversification. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 691–692, New York, NY, USA, 2006. ACM Press.

[6] T. Upstill, N. Craswell, and D. Hawking. Query-independent evidence in home page finding. *ACM Transactions on Information Systems (TOIS)*, 21(3):286–313, 2003.