# Case Western Reserve University at the TREC 2006 Enterprise Track

Adam D. Troy and Guo-Qiang Zhang
Department of Electrical Engineering and Computer Science
Case Western Reserve University
Cleveland, Ohio USA
{adam.troy,guo-qiang.zhang}@case.edu

## 1 Summary

For Case Western Reserve University's debut participation in TREC, we chose to participate in the Enterprise Track expert search task. Our motivation for participation stems from our work developing expert search capability for our prototype vertical digital library, MEMS World Online[1]. Our work incorporates two unique aspects. First, our relevance ranking mechanism relies on term position within each document rather than the number of term occurrences. This mechanism takes into account both term document rank and term co-occurrence proximity. Second, the expert score of closely related colleagues has a small effect on the score of each related expert. This follows the intuition that experts on a particular topic within a single organization tend to closely collaborate with one another. We also make some use of WordNet synonyms. We submitted a total of three runs to this years expert search task.

## 2 Relevance Ranking

For these experiments, we used a relevance ranking formula that incorporates the following aspects of the document set: term weights, taking into account term document frequency and estimated expected term document co-frequency across the document set, and the term occurrences, taking into account term rank and co-proximity in each document. Using term position to determine relevance is not a new idea [3, 1]. The weight mechanism assigns a weight for each term of the query as well as each pair of terms in the query. The goal of the weights, as usual, is to put emphasis on those terms in the query that are less common in the document collection. These terms have more power in differentiating which documents are most relevant to the query. Likewise, more emphasis is placed on those pairs of terms that are less likely to occur together. Unlike conventional approaches, the weights are not pre-computed as part of document indexing. Rather, each is computed on the fly since the weights vary depending on the combination of input terms. The weighting formulae are influenced by the Okapi weighting method [5].

$$\frac{S}{N} * log(\frac{2 * max_{df} + 0.5}{df + 0.5}) \tag{1}$$

Formula 1 shows the single term weight where $df$ is the number of the documents that the particular occurs in, and $max_{df}$ is equal to the greatest $df$ for the given query. The logarithm is used to dampen the range of the weights. $S$ is a constant which along with $N$, the number of terms, determines the total weight given to the single terms. Weights are also computed for each pair of terms, $i$ and $j$ in the query according to formula 2.

$$\frac{M}{\binom{N}{2}} * log(\frac{2 * max_{df} + 0.5}{df_i + 0.5} * \frac{2 * max_{df} + 0.5}{df_j + 0.5}) \tag{2}$$

[1] memsworldonline.case.edu

Similar to the previous formula, $M$ is a constant which along with $\binom{N}{2}$, the number of term pairs, determines the total weight for the term pairs. As stated above, this attempts to put emphasis on those pairs of terms which are least likely to occur together. Intuitively, $S$ and $M$ determine how much emphasis document term rank versus term co-proximity are given in the final relevance score. This correlation weight seems to be the first of its kind at least in this setting, a second-order information that takes into account query phrases in which several terms can occur at once.

For single term occurrences, the occurrence values are simply obtained as the inverse of the normalized distance of the term from the beginning of the document, shown in formaul 3 where $x_{ik}$ is the normalized position of term $i$ in document $k$. This function gives preference to terms appearing earlier in the document rather than later, or term document rank.

$$log(\frac{1}{x_{ik}}) \tag{3}$$

$$log(\frac{1}{|x_{ik} - x_{jk}|}) \tag{4}$$

For each pair of terms $i, j$, the (co-)occurrence value gives an indication of how closely the terms occur within each document, or term proximity, as shown in formula 4. These term rank and term proximity values are based on the intuition that the most important terms in a document will occur near the beginning of the document, which seems particularly true for email communications, and that terms which are part of a related concept will occur near one another. The dot product of the weight and occurrence terms is then computed to determine the relevance score for each document. The expert relevance score is then the summation of all the documents for which they are an author. In cases where the documents can be authored by more than one person, the document scores are weighted by the inverse of the number of authors, though this does not come into play because we use only emails for this work. Many more complicated and possibly more effective author scoring schemes are certainly possible.

A thorough analysis of these techniques for general relevance ranking will appear in the future. Preliminary experiments show that combining term co-proximity techniques with traditional ranking mechanisms such as pivoted length normalization [6] and Okapi [5] yield improved retrieval performance.

## 3 Collaborator Influence

One of the key aspects of an organization is the collaboration that takes place between it's members. We attempt to take into account the influence of this aspect in a small way. Essentially, if a potential expert has strong ties to any other high scoring experts, he or she receives a small increase in his or her score. This follows the intuition that experts in the same topic within the same organization have a strong likelihood of extensively collaborating together. Experts on common topics are likely to form closely connected clusters within the social network representing the organization of interest. Other work has used similar approaches [2]. This has also been inspired by our previous work studying scientific collaboration networks [7].

$$r' = r * (1 + C * (\frac{R}{r_{max}})) \tag{5}$$

$$R = \sum \frac{W * r}{n} \tag{6}$$

Equation 5 shows the calculation of a new author relevance score, $r'$ based on the original expert score $r$. $C$ is small constant, around 0.01 and $R$ is the collaboration strength weighted average of the score of the candidate experts collaborators in the social network constructed from the email corpus, as calculated in 6, where $W$ is the weight constant. This is divided by the maximum author score for the particular topic. Essentially, if an expert strongly collaborates with other high scoring collaborators they can receive up to a 1% increase in their score. This is a small effect because it should only come into play when two experts have very similar expert scores, the documents should still be the key component of the experts relevance. This technique was used for all runs.

# 4 Other Details

## Document Preprocessing

For the expert search task we relied solely on the email list portion of the W3C corpus. Document preprocessing was limited to HTML parsing, subject line extraction and author extraction. The subject lines were then used to identify email threads, which in turn were used to build a weighted social network of the organization. There are some considerations needed for the construction of the network. In order to somewhat limit the number of edges in the network, only two-way communication is taken to define a link in the network. For example, an individual sends an email to several people, only one of which responds. A link is only formed between the original author and the responder, not the other non-responsive recipients.

## Queries

For all submissions we used only automatic title queries. This was done for a few reasons. First, the full descriptions contain a great deal of extra terms which would likely have negatively affected performance. In order to use the descriptions effectively, advanced query analysis would be needed. For example, some descriptions describe what is *not* desired, using such terms naively would certainly not be useful. Secondly, automatic title queries are probably most similar to real user queries. Particularly when the system developer is formulating the queries, manual queries are not very similar to real user queries.

In retrospect, we would have liked to do one manual run, mainly for comparison purposes.

## WordNet

WordNet [4] was used to identify synonyms for one of the submitted runs. We did not expect great results through the use of WordNet because of the technical nature of most topics used this year.

## System Details

Two systems were used for these experiments, one implemented in python and the other in C++. Both use Berkeley DB databases. The python version allows for fast implementation of new ideas for testing on sample data sets while the C++ version allows for the development of an efficient system for larger volume runs.

# 5 Submitted Runs

We submitted a total of three runs for the expert search task. They are described below.

| Run | MAP | R-prec | B-pref | R-rank | P@10 |
|---|---|---|---|---|---|
| Emails | 0.2118 | 0.2757 | 0.2450 | 0.6615 | 0.5794 |
| Replies | 0.1910 | 0.2506 | 0.2246 | 0.6406 | 0.5434 |
| WordNet | 0.2154 | 0.2818 | 0.2523 | 0.6368 | 0.6116 |

Table 1: Expert search run result summary.

| Run | MAP | R-prec | B-pref | R-rank | P@10 |
|---|---|---|---|---|---|
| Emails | 0.0996 | 0.1479 | 0.1409 | 0.5233 | 0.3362 |
| Replies | 0.0778 | 0.1157 | 0.1108 | 0.4504 | 0.2615 |
| WordNet | 0.0831 | 0.1300 | 0.1232 | 0.4084 | 0.2752 |

Table 2: Supported expert run result summary.

- **Emails** — This run used the complete email list with the techniques described above, giving higher weight to reply emails.
- **Email Replies** — This run used only the email replies, as they are more likely to indicate expertise.
- **WordNet** — This uses the email list with WordNet synonyms.

A summary of results from each of the runs are shown in tables 1 and 2. Table 1 shows the expert search results without respect to support documents, while table 2 shows the results where relevant retrieved support documents are required. The most surprising result from this is the unexpected improvement of expert retrieval using WordNet synonyms. In the supported expert statistics on the other hand, WordNet generally resulted in poorer performance. The poorer performance of supported expert retrieval in general is not surprising, we made no attempt in retrieving all support documents, particularly in only using the email portion of the corpus. Also not surprisingly, using only reply emails yields lower performance.

# References

[1] M. Beigbeder and A. Mercier. An information retrieval model using the fuzzy proximity degree of term occurences. In *SAC '05: Proceedings of the 2005 ACM symposium on Applied computing*, pages 1018–1022. ACM Press, 2005.

[2] B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *DMKD '03: Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 42–48. ACM Press, 2003.

[3] E. M. Keen. Term position ranking: some new test results. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 66–76. ACM Press, 1992.

[4] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[5] S. E. Robertson, S. Walker, and M. Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track. In *NIST Special Publication 500-242: Proceedings of the Seventh Text REtrieval Conference (TREC-7)*, pages 253–264, July 1999.

[6] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM Press, 1996.

[7] A. D. Troy, G.-Q. Zhang, and M. Mehregany. Evolution of the Hilton Head Workshop research community. In *Education Digest of the 2006 Solid-State Sensor and Actuator Workshop*, June 2006.