# CSIR at TREC 2007 Expert Search Task

Jiepu Jiang, Wei Lu, Dan Liu
Center for Studies of Information Resources, School of Information Management
Wuhan University, P. R. China
{jiangjiepu@gmail.com, reedwhu@gmail.com, liudan1987@gmail.com}

**Abstract.** This is the second year for the participation of Center for Studies of Information Resources (CSIR) in the TREC Expert Search Task. Rather than using the candidate profile based approach, a simplified two stage approach is used in our experiment, that is, documents are ranked based on topics, and each expert is scored intuitively by the weights of documents the expert appeared. Instead of the modeling of expert search, we have mainly focused on the effect of document filtering in the expert search. In our experiment, only the top *n* ranked topic-relevant documents where the expert also appeared are calculated into the expert score. The tuned value of *n* under W3C corpus for a best performance is 10, which is proved to be stable under CERC corpus.

## 1. Introduction

The Expert Search Task of TREC Enterprise Track, quite different from the traditional ones, aims at creating a ranked list of experts in given topics rather than relevant documents. In TREC 2005 and TREC 2006, a canonical list of people, namely a list of names and email addressed of the candidate experts, was extracted officially and provided to all participants along with the corpus data. A ready-made candidate list lessened the procedure of expert recognition and helped the participants to concentrate on the IR models.

TREC 2007 has chosen CSIRO as the enterprise of interest and proposed some new challenges for participants. Most obviously, the candidate list is not available anymore and participants need to extract experts themselves. Though a real situation it is when searching for experts from enterprise website, the absence of candidate list has brought the participants a series of problems. Firstly, the recognition of person names and email addresses is required; secondly, it is necessary to filter out persons appeared in the corpus because they may be ones outside the CSIRO or stuffs rather than experts even inside the CSIRO; furthermore, the occurrences of an expert in the corpus are diverse in format and a mergence is needed accordingly.

The previous works of TREC Expert Search Task, reflected by the notebooks of the participants, have the following two characteristics. On the one hand, the expert profile creating approach was widely used and implemented diversely. Among them, the window-based technique, which is included in systems of 6 participants [1, 2, 3, 4, 5, 6] out of 9 who used the expert profile creating approach [1, 2, 3, 4, 5, 6, 7, 8, 9], is the most frequently adopted method. Besides, Jennifer et al [1] proposed the top sentence approach, the whole document approach and the summarization approach to create profiles of experts. In their work, different methods of profile creating are performed separately and merged into the final profile of experts. On the other hand, the two stage approach of expert search is also frequently used and receives good performance. Besides, external information was used by some participants and proved effective to improve the final result. For example, Zhu et al [2] considered the page rank of each document as the way to measure the importance of each document. Jennifer et al [1] utilized the Google Scholar to return publications of each candidate and treated them as supplementary information to judge the expertise of candidate.

Last year, we used a window-based approach to create expert profile and then retrieved the profile information to generate a ranking list of experts by traditional IR method. This year, we reverted to the two stage approach which is simplified and focus mainly on the effect of document filtering in expert search. In our experiments, documents are ranked based on topic and only the top ranked topic-relevant documents where the expert appeared are summed into the expert score. It is assumed in our experiment that the expert who occurs in the top ranked documents to a given topic should be more relevant to the topic, which is similar to Thijs Westerveld [7]. By using this method, a better result is achieved this year. Due to the time limitation, external information and the query expansion approach are not used in our experiment, although they are generally accepted as an effective way of enhancing the final results.

In section 2, an emphasis on the approach and model adopted for this year is given. In section 3, we give a brief introduction to the whole procedure of our experiment. The analysis and evaluation of the results are discussed in section 4. At last, a conclusion is made and the future work is pointed out.

## 2. The Approach and the Model

There are usually two approaches of expert search: the expert profile creating approach and the two stage approach. The former, as mentioned in the introduction, collects evidence around the expert occurrence as the profile for experts and used the profile to match the query. However, the combination of document fragments into the profile of an expert may change the original semantic meaning of the documents. For example, given a topic “semantic web” and three documents, the first document contains one phrase “semantic web”; the second document contains one “semantic” in the phrase “semantic translation” and the third document contains two “web” in the phrase “web mining”. Obviously, the first document is relevant while the other two not. But if the second and the third document are combined together, the traditional IR model will treat the new document as one more relevant to topic “semantic web” than the first one, although it is not in fact. To avoid this problem, we have adopted the latter approach. This approach assumes that the expert appears in the relevant documents of a given topic possibly possesses expertise of this area. As a result, traditional IR method is used firstly to generate ranked lists of the relevant documents to each given topic. Persons occur in the relevant documents will be considered later as the possible experts of the topic.

Given a query $Q$, a result ranking of documents is retrieved by traditional IR method and $W_i$ is the weight for the $i$th document retrieved which is computed based on BM25. Then $W_e$, the weight of the expert $e$ for query $Q$, equals to the linear combination of the weight of documents which are retrieved for $Q$ and contain occurrences of the expert $e$’s evidence. The simple formula is as follows:

$$w_e = \sum_{i \in D} w_i \qquad (1)$$

*Where D is the set of relevant documents for query Q containing the occurrences of the expert e.*

In the document ranking list, the higher the document ranks, the more relevant it is to a given topic. On the contrary, it is quite likely that the documents with lower ranks are not relevant to the topic but will affect the effectiveness of retrieval. Thus, we assume that a cutoff of the document list should be useful for filtering the irrelevant documents. The cutoff value is tuned based on data of TREC 2006 and the best tuned value is used for our runs this year. Details of our experiment will be discussed in section 3.

Besides using the ranking list according to topic relevance, we also use a document ranking list by expert evidence (expert name and expert email). That is, we firstly get a document ranking list by expert evidence, and then use the topics for filtering. The underlying idea is that the document which contains more evidence of an expert should be more useful in representing the expert’s expertise. As the above one, a cutoff is also set for this method in our experiment.

# 3. Experiment

Our experiment is based on OKAPI 2.51 in a Linux environment. The whole procedure can be divided into 3 steps.

## 3.1 Data Cleaning

The CERC collection contains various resource crawled from CSIRO website. Most of the documents in the collection are html pages, while other resource such as formatted files, plain texts, octet-stream files, script, logs etc. are also included. A glance at the ingredient of CERC collection is in Table 1.

| **Data Type** | **Details** | **Document Count** | **Percentage** |
|---|---|---|---|
| HTML Page | text/html | 328546 | 88.62% |
| Formatted File | pdf, word, rtf, ppt, excel | 13682 | 3.69% |
| Octet-stream | application/octet-stream | 10080 | 2.27% |
| Unknown data | unknown | 10000 | 2.70% |
| Plain Text | mostly log file | 6271 | 1.69% |
| Script | CSS, JavaScript | 957 | 0.26% |
| Multimedia | asf, real, wmv, bmp | 472 | 0.13% |
| XML | text/xml | 179 | 0.05% |
| Other |  | 528 | 0.14% |
| Total |  | 370715 |  |

Table 1: The ingredient of CERC collection.

Html tags, semantic or non-semantic, impede text retrieval to a certain extent. The non-semantic html tags, mostly used only for controlling the display of web pages, provide little information for retrieval and need to be removed. Content within semantic html tags is usually more important and should be assigned to a higher weight at retrieval. We use html-parser [10] to traverse the pages and output cleaned corpus data. During the process of traverse, content within <title> and <meta> is extracted into separate field, while other tags are removed except for a part of <a> tags which are linked to email addresses because they are useful in expert recognition.

Formatted files, when returned after an http request, are described in self-defined XML tags. As a result, the same method which is used on html pages can be applied to clean data and extract important information. Furthermore, script and multimedia resource is dismissed since it contains little information of expert and contributes no more than 0.5% of the corpus data. Other resource is retained for the difficulty of cleaning due to diversified formats.

## 3.2 Expert Recognition

The absence of candidate list this year requires a task of expert recognition. Though person names and email addresses can be extracted from documents, filtering is needed to distinguish experts from staffs, experts inside the enterprise from those outside. Further, a combination of results is required to merge the email address and different name formats of the same expert.

The domain of email address is an indication direct and effective for distinguishing experts inside from outside. On the assumption that each expert owns an email at the domain of the enterprise and the email appears at least once in the documents, the procedure of expert recognition in the system includes the following three steps.

Firstly, email address is extracted from documents and divided into two groups by judging whether its domain belongs to the enterprise or not. The group of emails at the enterprise domain will be used in the next step to generate characteristics of each expert. In this group, the email with a username containing words seldom used for person name is filtered for its high possibility to be and email for public use. Another group, supposed to be mostly owned by outsiders, will provide a determinant to avoid a person name from being recognized as a valid expert if hardly any positive clew exists. The documents where email occurs are recorded and merged into the occurrences of experts in the next step.

In the second step, an automatic procedure is performed to generate characteristics of experts and record the documents where they appeared. The characteristic of an expert is defined as a triad, <email, name, occurrence>, and can be renewed by any representation which is compatible and owns information more specific. Two sets of characteristics can be created in the system to represent groups of expert inside and outside the enterprise, namely $E_i$ and $E_o$. The whole procedure can be described as follows: for each priority, $E_i$ and $E_o$ are scanned orderly with rules of this priority to renew every element within them, that is, if a name found in the document accords with current rule and does not exist in the other group, it can be absorbed into the current element with the occurrence recorded.

In the last step, each expert is identified in the designated format by the most specific person name found in the documents.

Though no official list of expert recognition is available in this task, it is revealed in the official list of relevant experts that 29 experts out of 152 who appeared in the relevant list were not found in our experiment. Besides the flaw of techniques in expert recognition process, the reason may reside in the incomplete truth of the undertaken assumption.

| Priority | Rule |
|---|---|
| 1 | Person name with hyper-link to an email address |
| 2 | Full name absorbable appeared in the same page |
| 3 | First name is initial and absorbable, appeared in the same page |
| 4 | Person name appeared in the same page |

Table 2: Priority and the according rules in expert recognition.

## 3.3 Expert Retrieval

When experts are recognized and their occurrences are found in documents, traditional IR methods are used to retrieve relevant documents of a given topic. In our experiment, we use OKAPI and model BM25 to compute the relevance score of documents and generate a ranking list of relevant documents to each topic.

As mentioned in Section 2, the weight of an expert to a topic equals to the linear combination of the weight of documents which are retrieved for the topic and have the occurrences of the expert evidence. But the documents with a lower rank are highly possibly not relevant to the given topic, which means if an expert occurs in a large amount of the lower ranked documents, the linear combination of the weight of documents will largely increase the expert weight, although they cannot truly support the expert to be relevant with the given topic indeed.

To solve this problem, a cutoff *n* is set to filter out the lower ranked documents, that is, only the top *n* ranked topic-relevant documents where the expert also occurs are located to compute the relevance score for the expert. As illustrated in Table 2, this filtering is effective for the final result.

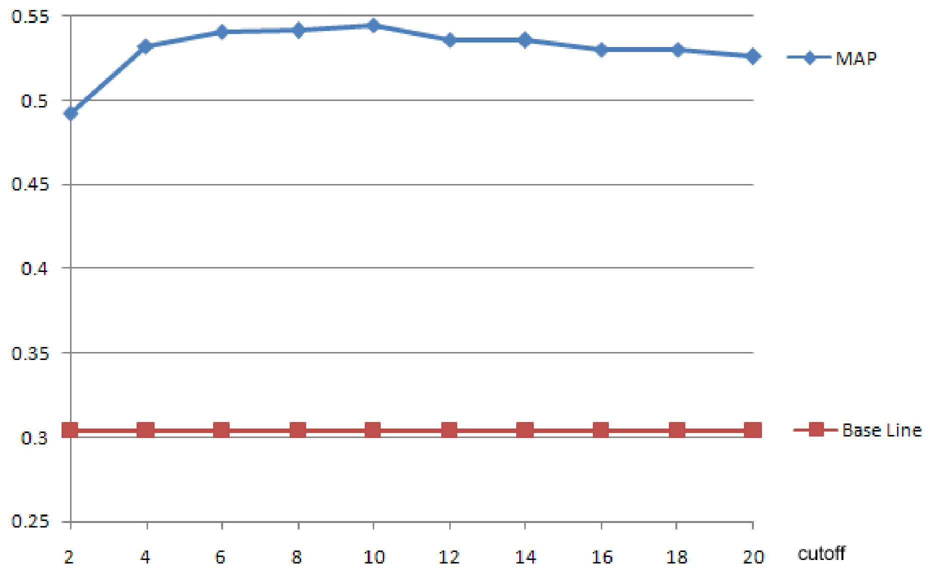

Figure 1. MAP results at different cutoff values, using data of TREC 2006

Figure 1 shows the MAP results based on the experiments conducted under different cutoff *n*. A baseline result of the experiment which involves no filter of ranked documents is given to make a contrast. It is revealed evidently that a great improvement has been made when filtering is considered in the experiment. MAP keeps increasing along with the cutoff *n* until it reaches the peak when cutoff *n* is set to 10. Then MAP begins to decrease gradually.

The already trained cutoff value 10 is used directly in our run WHU10 this year and receives the best MAP results of our 4 runs as we expected before. WHUC5 and WHU15, where the cutoff value is assigned to 5 and 15, are also submitted to be the control group. Another run, WHUE10, is conducted under a different method, that is, we firstly get a document ranking list by expert evidence (person name and email), and then use the topics for filtering. In this run, cutoff is also set to 10. The results of each run will be emphasized in section 4.

## 4. Evaluation

In section 2 and section 3, the basic model and method adopted and the whole procedure of our experiment are introduced. This year, we submitted 4 runs to TREC for evaluation. Table 3 gives a general look at the evaluation of these 4 runs and a baseline run.

| **Run** | **cutoff** | **MAP** | **R-PREC** | **BPREF** | **RECIP-RANK** | **P@10** |
|---|---|---|---|---|---|---|
| baseline | -- | 0.2582 | 0.2030 | 0.5934 | 0.3866 | 0.0980 |
| WHU10 | 10 | 0.3399 | 0.2862 | 0.6392 | 0.4738 | 0.1220 |
| WHUC5 | 5 | 0.2193 | 0.1483 | 0.5579 | 0.3054 | 0.0620 |
| WHU15 | 15 | 0.3060 | 0.2744 | 0.6392 | 0.4404 | 0.1120 |
| WHUE10 | 10 | 0.3280 | 0.3105 | 0.6399 | 0.4674 | 0.1020 |

Table 3. Results of our 4 run this year and the baseline results.

WHU10 uses the best value of cutoff trained for data of TREC 2006 and receives the best MAP result in our 4 runs. WHUC5 and WHU15 use the 5 and 15 as the cutoff for filtering and receive relatively worse performance than WHU10. The baseline run is conducted without considering cutoff for filtering and receives worse result than WHU10 and WHU15. WHUE10 is conducted under a different method as we mentioned in section 3 and receives results a little lower than WHU10 in MAP, RECIP-RANK and P@10 but better in R-PREC and BPREF.

Results of WHU10 and WHU15 and the contrast between them and the baseline accord with the outcome of experiment under the data of TREC 2006. The method used on WHUE10 is also proved to be an effective way for expert search. Though no experiment is achieved to find the best cutoff value of WHUE10, it receives results better than WHUC5 and WHU10 and a little worse than WHU10.

## 5. Conclusion

In the TREC Expert Search Task this year, a better result, respectively, is achieved by our system than the work of last year. The future work will mainly concern with refinements of our model.

Firstly, the frequency of an expert appeared in one document is not considered in the method used for WHU10, WHUC5 and WHU15, which could also be taken into account in the computation of document weight. Further, a thorough experiment is needed to find out the best cutoff value in the method used for WHUE10 and to make a judgment between the different two methods of the document ranking based approach.

## References

[1] Jennifer Chu-Carroll, Guillermo Averboch, Pablo Duboue, David Gondek, J William Murdock, John Prager, Paul Hoffmann, Janyce Wiebe. IBM in TREC 2006 Enterprise Tack. In the *Proceedings of the 15th Text Retrieval Conference*, 2006.

[2] Jianhan Zhu, Dawei Song, Stefan Rüger, Marc Eisenstadt, Enrico Motta. The Open University at TREC 2006 Enterprise Track Expert Search Task. In the *Proceedings of the 15th Text Retrieval Conference*, 2006.

[3] Zhao Ru, Qian Li, Weiran Xu, Jun Guo. BUPT at TREC 2006: Enterprise Track. *Proceedings of the 15th Text Retrieval Conference*, 2006.

[4] Wei Lu, Stephen Robertson, Andrew Macfarlane, Haozhen Zhao. Window-based Enterprise Expert Search. *Proceedings of the 15th Text Retrieval Conference*, 2006.

[5] Ganmei You, Yaojie Lu, Gang Li, Yueyan Yin. Ricoh Research at TREC 2006: Enterprise Track. *Proceedings of the 15th Text Retrieval Conference*, 2006.

[6] Shenghua Bao, Huizhong Duan, Qi Zhou, Miao Xiong, Yunbo Cao, Yong Yu. Research on Expert Search at Enterprise Track of TREC 2006. *Proceedings of the 15th Text Retrieval Conference*, 2006.

[7] Thijs Westerveld. Correlating Topic Rankings and Person Rankings to Find Experts. *Proceedings of the 15th Text Retrieval Conference*, 2006.

[8] Krisztian Balog, Edgar Meij, Maarten de Rijke. Language Models for Enterprise Search: Query Expansion and Combination of Evidence. *Proceedings of the 15th Text Retrieval Conference*, 2006.

[9] Desislava Petkova, W. Bruce Croft. UMass at TREC 2006: Enterprise Track. *Proceedings of the 15th Text Retrieval Conference,* 2006.

[10] http://htmlparser.sourceforge.net/