Research on Enterprise Track of TREC 2007

Huawei Shen, Guoyao Chen, Haiqiang Chen, Yue Liu, Xueqi Cheng

Institute of Computing Technology

Chinese Academy of Sciences

ABSTRACT

We (ICT-CAS team) participated in the Enterprise Track of TREC 2007. This paper reports our experimental results on this track.

1. INTRODUCTION

The goal of the Enterprise Track is to study the issues that arise when searching the documents of an enterprise (organization). It involves some new things, including new data and new tasks.

There are a variety of document types: web pages, news/email archives, document archives and many other document types. It is critical for enterprise track to utilize the particular characteristics of each document type appropriately.

Different from Enterprise Track 2005 and 2006, this year's enterprise track is explicitly divided into two tasks, document search task and expert search task. What's more, a new corpus is used instead of the W3C corpus adopted in the last two years' enterprise track. This new corpus is provided by CSIRO.

In last two year enterprise track, several models have been proposed to combat with expert search task. Among them, two-stage language model [1] has been proved to be a successful model for expert search task. And many systems have been implemented with this model. This model consists of a document relevance model and a co-occurrence model. The document relevance model retrieves documents which are relevant to the query associated with expertise topic. Then, the co-occurrence model is used to find experts who are closely related to the expertise topic. Here, document is taken as a hidden variable, separating the query from the candidate experts.

This year, our team's system adopts the framework of the two-stage language model. Document search task and expert search task are addressed in this uniform framework. Firstly, documents are scored and ranked using BM25 retrieval model. Secondly, occurrence of each candidate expert is recognized and the score of each candidate expert is the aggregation of the score of all documents, which are relevant to the given query and contains more than one occurrence of this candidate expert.

The score for each document, produced by BM25 retrieval model, is used to rank the documents. And the ranked list of documents is submitted as the result of document search task. As for each topic, up to 1000 relevant documents are retrieved.

As to the expert search task, a ranked list of experts is submitted as the result for a given topic. Each list contains no more than 100 experts. The list is sorted according to the score of each expert.

This report is organized as follows. Section 2 introduces the new things occurred in this year enterprise track. We discuss the query expansion and formulation in Section 3. Document search task are discussed in Section 4. Section 5 describes the expert search task. Section 6 concludes this report.

2. What's New in Enterprise Track 2007

Different from the last two year, there are two tasks in this year enterprise track. As stated in last section, document search task is explicitly divided from expert search task. Investigating whether the result of document search affects the result of expert search is an interesting problem.

CERC [2], CSIRO Enterprise Research Collection, is introduced as a new enterprise corpus. This corpus is a document collection for the 2007 track, which is a crawl of the publicly available web pages from *.csiro.au domain. The total size of this corpus is 4.18 GB. The document collection consists of 370715 documents, stored in 267 bundles. The corpus has a mixture of document types, including web pages, document archives (.pdf, .doc, .ps, .xls, .bib, .rtf, .tex and many other formats), java scripts, etc.

In addition, the topics are all from real users, including examples of "key pages" that should be retrieved. The judgments are also from real users.

As to expert search, no pre-defined list of candidate experts is provided. And for each topic, there are only a few experts (typically one or two).

3. Query Expansion and Formulation

A searcher's request on a topic is often complex and needs to be converted to queries for IR systems to process. For example, if topic is "sustainable agriculture", documents which are about "sustainable environment" are not so relevant. So, the query expansion and formulation is very important for IR systems.

A topic, provided by the organizer of enterprise track, consists of three fields, including title, narrative, key pages. The title field of a topic is typically used as the search query. However, by considering the narrative field, this query can often be expanded to form a number of more precise and informative queries leading to better search results. For an example, the title of the topic CE-001 is "Genetic Modification". If only the title field is used to form the query, the documents, which discuss biotechnology or GM and do not contain "genetic modification", are not retrieved. Examining the narrative field of this topic, we find that there are several other informative words or word pairs, including "gene technology", "biotechnology" and "GM".

The "key pages" field is for feedback runs and all our submitted runs are not feedback runs. So, we only use the title and narrative field of the topic to form the query.

The process of query expansion in our system is automatically completed. We simply delete all the stopwords in the narrative field and the remaining content are divided into several fragments. All this fragments are used to expand the query.

4. Document Search Task

For document search, systems will return ranked lists of documents from the CERC collection. Retrieved documents should tend to be authoritative pages such as project homepages and documents dedicated to the given topic, rather than pages that make passing mention of the topic.

The anchor texts on the hyperlink to a document are considered as the content of this document. In addition, the keywords in the meta field of a webpage (document) are also considered as the content of this document.

For our system, we adopt BM25 retrieval model, a well-known probabilistic model [3].

Assume d is document belonging to the collection. We regard it as a vector $d = (d_1, \cdots, d_V)$, where d_j denotes the term frequency of the j th term in d and V is the total number of terms in the vocabulary. The score of document d is computed by

$$score(d) = \frac{(k_1+1)d_j}{k_1((1-b)+b\frac{dl}{avdl})+d_j}\log\frac{N-df_j+0.5}{df_j+0.5}, \qquad (1)$$

where df_j is the document frequency of term j, dl is the document length, $avdl$ is the average document length across the collection, and k_1 and b are free parameters. The optimal values for k_1 and b in CERC are respectively 1.5 and 0.3.

Our team submits 4 runs for document search task. Table 1 describes the results of our four runs for document search task. We can see that the two runs based on BM25 retrieval model are better than the other two runs.

Table 1. Results of each run for document search task

Runs	MAP	MRR	Bpref	Recall@1000
DocRun01	0.3970	**0.8384**	0.3963	**0.8580**
DocRun02	**0.4048**	0.8158	**0.4013**	0.8322
DocRun03	0.1746	0.3493	0.2276	0.8322
DocRun04	0.3316	0.8014	0.3263	0.8322

The first and the second runs are based on BM25 retrieval model. The only difference is the query for each topic. The query for the first run are formed using only the title field of the topic, however, the query for the second run are based on both the title field and the narrative field of the topic.

In addition, we try to use PageRank [4] algorithm to re-rank the resulting ranked list of the second run. We choose the top 200 documents in the ranked list for each topic and construct directed graph, where the nodes are documents (web-pages) and the arcs are hyperlinks between them. We apply PageRank algorithm to the obtained graph with the parameter c=0.85. By re-ranking the top-200 documents, we obtain a new ranked list of documents for each topic. This is our third run for document search task.

The forth run are based on the second and the third runs. The score of the document in the forth run are based on the positions of this document in the second and third runs.

5. Expert Search Task

5.1 Expert Identity Recognition

In this year, there is not a list of candidate experts. We need find all the expert identities occurred in the document collection. Email is regarded as the unique identity of an expert.

All the valid email addresses are in the format: firstname.lastname@csiro.au. However, several other email addresses should also be considered valid because of there are several sub-domains under the domain csiro.au, such as atnf.csiro.au, ento.csiro.au, cse.csiro.au and many other sub-domains. If the username (the part before @ in the email address) of one email address is the same as that of another email address, these two email addresses are regarded being associated with the same person. For example, julie.carter@csiro.au, julie.carter@dwe.csiro.au and julie.carter@ento.csiro.au are associated with the same person. We try to find all the possible valid email addresses. We find all the occurrences of the symbol '@' and then judge whether this occurrence of symbol '@' is the component of an email address. By this method, we get lots of candidate email address, and then we filter the invalid email address. Finally, by incorporating the email addresses which have the same username, we get all the email addresses of candidate experts.

The real users in CSIRO provide total 152 email addresses as experts for all the 50 topics. Our expert identity recognition method recalls 129 of the 152 email addresses. The recall rate is about 85%.

After having got all the valid email addresses of candidate experts, we get the full name of candidate expert from her/his email addresses. For the example in last paragraph, the possible name for these email addresses is "Julie Carter". Other variety of expert names is also needed to be considered to improve the recall of the recognition. We use an automatic method to generate such variety of expert names. For the example used in this section, variety of names may include "J. Carter", "Ms. Carter", "Dr. Carter", "Prof. Carter", etc.

Up to now, we have got all the identifiers for each candidate experts, including email addresses, variety of expert names. We pre-process all the documents in the collection by replacing all the non-ASCII characters with spaces, removing HTML mark-up in web pages, replacing sequential spaces with a single space. Then we use the Aho-Corasick algorithm [5] to match these expert identifiers against the pre-processed documents.

5.2 Expert Search

For expert search, systems are required to return ranked lists of email addresses representing person. Our team submits 4 runs for expert search task. Table 2 shows the results for each run. The second run using two-stage language model is the best one. The forth run using the new method presented by us is in the second place.

Table 2. Results for each run for expert search task

Runs	MAP	MRR	Bpref	Recall@100
ExpertRun01	0.3066	0.4695	0.6784	0.6579
ExpertRun02	**0.3689**	**0.5142**	0.6851	**0.6645**
ExpertRun03	0.0146	0.0140	0.6844	0.6513
ExpertRun04	0.3433	0.5077	**0.6884**	0.6645

In the framework of two-stage language model, the score of expert is based on the score of documents in which the expert is mentioned. When we have obtained the score of all the documents given a query of a topic, the score of expert e can be calculated by

$$score(e) = \sum_{d \in D} score(d) \times NumberOfOccurrence(e,d) , \qquad (2)$$

where D is the set of documents in the collection CERC, and NumberOfOccurrence(e, d) denotes how many times expert e is mentioned in document d .

The first and second run for expert search task are respectively based on the first and second run for document search task. The experts' score in these two runs are calculated by (2).

As to the third run for expert search, we construct a profile for each candidate expert [6]. The profile of an expert is simply the concatenation of the documents in which this expert is mentioned. Then the rank of the expert is based on the score the corresponding profile. Given a query of a topic, the score of the profile are computed by BM25 retrieval model.

For the forth run, we try a new method. We construct a document-expert graph from the mention relation between documents and experts. Given a query of a topic, we find all the relevant documents and construct a sub-graph from the complete document-expert graph. Then using HITS algorithm [7], each document is assigned a hub-value and each expert is assigned a authority-value. Finally, a ranked list of experts is obtained according to the authority-value for each expert.

6. Conclusions and Future Work

This paper reports the experiments of our team on Enterprise Track 2007. We mainly design an expert search system based on the two-stage language model. The run based on two-stage language model is the best one in our four runs submitted. In addition, we try a new method based on HITS algorithm and the result is slightly worse than the run using two-stage language model.

Most existing models focus on the relevance of expert given a topic. Only little attention is paid to the authority of the expert. Intuitively, the prior authority can be used to improve the effectiveness of existing models. Our team intends to explore how to obtain the knowledge of authority of expert as our future work.

7. ACKNOWLEDGMENTS

We thank to the member of social computing and P2P group in the Institute of Computing Technology, Chinese Academy of Sciences.

8. REFERENCES

[1] Cao, Y., Liu, J., Bao, S. and Li, H. Research on Expert Search at Enterprise Track of TREC2005. In Proc. Of TREC2005.

[2] http://es.csiro.au/cerc/

[3] S.E. Robertson, H. Zaragoza, M. Taylor. Simple BM25 Extension to Multiple Weighted Fields. In Thirteenth Conference on Information and Knowledge Management (CIKM), 2004.

[4] S. Brin, L. Page, R. Motwami, T. Winograd. The PageRank citation ranking: bring order to the web. Technical Report, Stanford University.

[5] A. V. Aho and M. J. Corasick. Efficient string matching: An aid to bibliographic search. Communications of the ACM 18 (6): 333-340, 1975.

[6] Y. Fu, W. Yu, Y. Li, Y. Liu, M. Zhang and S. Ma. THUIR at TREC 2005: Enterprise Track. In Proc. Of TREC 2005.

[7] J. Kleinberg. Authoritative sources in a hyperlinked environment. In Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998.