Research on Enterprise track of TREC 2008

Huawei Shen, Lei Wang, Wenjing Bi, Yue Liu, Xueqi Cheng

Institute of Computing Technology

Chinese Academy of Sciences

ABSTRACT

This is the third year that we (ICT-CAS team) participated in the Enterprise Track of TREC. The track of this year includes two tasks, being document search task and expert search task. As to the document search task, we compare two retrieval models, namely BM25 model and language model. As to the expert search task, we focus on the authority of persons by exploring the persons' recommendation network constructed from their associated documents. What's more, we investigate the effectiveness of query expansion on both tasks.

1. INTRODUCTION

Using the same corpus (.CERC) [1] provided by CSIRO, the enterprise Track of this year is much alike to that of last year. However, the scenario is very different. As to last year, systems are designed to help the science communicators create an overview page in the given topic area. This page may include hyperlinks to some authoritative documents and several persons with desired expertise. The topics are provided by these science communicators. As to this year, systems are expected to help the enquiries staffers answer the questions from the users of CSIRO Enquiries. The topics are from CSIRO Enquiries. Thus, these topics are usually more specific. For several cases, the want is only a picture or a web page or even an email address for some person.

The document search task is similar to the traditional ad-hoc document retrieval task. Two models, BM25 model [2] and language model [3], are widely used and demonstrated to be effective. We compare these two models by experiments on the .CERC corpus for the document search task. The language model outperforms the BM25 model slightly.

In addition, query expansion is introduced. As known, the document search task pays more attention to the precision than recall. Only the highly-relevant authoritative documents are considered as the required results. However, query expansion [5] is biased due to topic drift while improving the recall performance. We investigate the effectiveness of query expansion by experiments and the results show that it is promising.

As to the expert search task, we focus on the authority of persons. Firstly, we make a difference between different types of occurrence of persons. Full name, short name, and email address are three main types of occurrence. Then, we construct a recommendation network between persons. Link analysis [7][8] is carried on this network to identify the authoritative persons.

The report is organized as follows. Section 2 describes the experiments on the document search task. Section 3 introduces our attempts to explore the authority of persons. Conclusion is given in section 4.

2. Document Search Task

Last year, BM25 model is employed in the document search task [6]. This year, we adopt the widely-used language model instead and compare these two models. Two reasons motivate us to prefer the language model. One is its foundation in statistical theory. Another is that it performs quite well empirically on ad hoc document retrieval.

Instead of addressing relevance directly, the language model approach asks a different question: *How likely is it that document d would produce the query q*? Firstly, a language model is estimated for each document. Then documents are ranked by the likelihood of the query according to the language model.

As to language model approach, smoothing is required. The main purpose of smoothing is to assign a non-zero probability to the un-seen words by discounting the probabilities of the words seen in the document according to the collection language model. Several smoothing methods have been introduced into language model approach. We employ the Bayes smoothing method due to that this method takes the length of document into account and is appropriate to document collection with heterogeneous length, such as the .CERC used in the Enterprise Track.

For each document, five fields are considered, including URL, title, keywords, anchor text and body text. The content of some field, such as keywords, can be empty. Each field is assigned a weight indicating its importance to retrieval task. In our system, the weights are tuned to 3, 1, 1, 3, and 1 respectively.

As shown in table 1, the language model always outperforms the BM25 model slightly no matter whether the query expansion is performed. Both models are configured with the best-tuned parameters.

Query expansion is a technology to match additional documents by expanding the original search query. However, this improvement of recall comes at the expense of reducing the precision. A tradeoff between recall and precision is needed. Actually, a rigorous query expansion can usually improve both the recall and the precision.

People can use pseudo-relevant documents or other resources (such as Wikipedia [10]) as the basis of query expansion. For example, the anchor texts in the article titled by the query are used to expand the original query. In our system, we rely on the pseudo-relevant documents, which are documents returned by our search engine using the original query. Usually, only the top-ranked documents are considered to be pseudo-relevant documents. The terms occurred in the pseudo-relevant documents are weighted and ranked. In our work, we choose two kinds of term weighting models. The one, known as Bo1, is based on Bose-Einstein statistics. The other is based on the Kullback Leibler (KL) divergence [1] between the pseudo-relevant documents and the collection. People can select the top-ranked terms as candidate for query expansion. We use the default setting of the parameters for query expansion, and choose the top 10 terms from the top 3 ranked documents, suggested by Amati in [5].

From table 1, we can see that query expansion can improve the precision and the recall for both the BM25 model and the

language model. We also can see that the KL method outperforms the Bo1 method.

Table 1. Results for the document search task

Methods	MAP	MRR	P@20
BM25 model	0.4534	0.8690	**0.6820**
Language model	0.4445	**0.8768**	0.6580
BM25 + Bo1	0.4666	0.8629	0.6450
LM + Bo1	0.4571	0.8728	0.6420
BM25 + KL	**0.4699**	0.8729	0.6660
LM + KL	0.4632	0.8630	0.6570

3. Expert Search Task

3.1 Document-person Association

Before expert retrieval, we need identify the occurrence of persons and further build the document-person association.

Person identification is an important research problem and has attracted many attentions. Several promising tools exist, such as Identifinder. Most of them are not free. Thus, we take another method [6]. Firstly, we obtain a list of candidates and then we identify their occurrences in corpus using Aho-Corasick algorithm [9]. Candidates are identified by their email addresses. Then, the full names or short names are generated according their email addresses. In addition, we expand our candidate list by using the candidate list provided by Krisztian Balog [3]. The final candidate list contains 3624 candidates.

Now we estimate the extent to which document d characterizes candidate ca, that is $p(d \mid ca)$. According to Bayes' Theorem,

$$p(d \mid ca) = \frac{p(ca \mid d) \cdot p(d)}{p(ca)}.$$

We assume $p(d)$ and $p(a)$ follow uniform distributions.

We make a difference between different types of occurrence of persons, including full name, short name, and email address. They are assigned with the weights being 7, 3 and 1 respectively.

All the types of occurrence of candidate are replaced by their unique identifiers. Different types are replaced by different number of the unique identifiers. For example, a occurrence of full name is replaced by seven times of its corresponding unique identifiers. Then, we treat the unique identifiers as a term and using the language model to estimate.

3.2 Authority of persons

Existing expert retrieval models mainly focus on the relevance between the query and the candidate person. The hypothesis behind these models is that the candidate persons with desired expertise usually occur much more times in the documents relevant to the query. In most cases, this hypothesis makes sense. However, it is not sufficient in some situation. For example, some persons with high popularity in the relevant documents are not the most desired persons. Thus, we need another dimension, called authority by us, to characterize the expertise of a person.

We construct a recommendation network of persons. If two persons occur in the same document and one person take the form of email address and another is a full name or short name. The first person is considered recommend the second person, and correspondingly a directed edge from the first person to the second person is added to the recommendation network. Then, we apply the PageRank algorithm [8] on this network to score the authority of each person. To evaluate the effectiveness of this method, the answers are set to be the persons who have a homepage under http://www.csiro.au/people/. The results show that our method has a high precision. However, the recall is low. This can be contributed to the sparseness of the network.

We aggregate the authority rank of person and the relevance rank of person into a final rank.

4. Conclusion and Future work

This paper reports the experiments of our team on Enterprise Track 2008. In document search task, we focus on query expansion based on two retrieval models, the BM25 model and the language model. In expert search task, we adopt two-stage language model based on the results of the former task, and pay much attention to the authority of persons.

In the future, we will devote to explore the authority of persons and investigate how it can be used to improve the expert retrieval.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

[1] CERC. http://es.csiro.au/cerc/.

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

[3] K. Balog. People Search in the Enterprise. PhD thesis. University of Amsterdam. June 2008.

[4] G. Amati. Probabilistic Models for Information Retrieval based on Divergence from Randomness. PhD thesis, Univ.of Glasgow, 2003.

[5] C. Macdonald and I. Ounis. Expertise Drift and Query Expansion in Expert Search. CIKM '07, November 6-8,2007.Lisboa, Portugal.

[6] H. Shen, G. Chen, H. Chen, Y. Liu, X. Cheng, Research on Enterprise Track of EREC 2007. In TREC 2007.

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

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

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

[10] Wikipedia, the free encyclopedia. http://en.wikipedia.org/.