CSIR at TREC 2008 Expert Search Task: Modeling Expert Evidence in Expert Search

Jiepu Jiang[1], Wei Lu[1], Haozhen Zhao[2]

[1] Center for Studies of Information Resources,
School of Information Management, Wuhan University
{jiepu.jiang, reedwhu}@gmail.com

[2] College of Information Science and Techonology, Drexel University
{haozhen.zhao}@ischool.drexel.edu

Abstract. In this paper, we described our method for the expert search task in TREC 2008. First, we proposed an adaption to the language modeling method for expert search, which considers the probability of query generation separately using each kind of expert evidence (full name, abbreviated name, and email address). Current expert search models can be easily integrated into our method. Our experiments indicated that our method can make use of the ambiguous evidence in expert search (abbreviated name), which often casued a drop in effects in other methods. Besides, we also used a probabilistic measure to detect phrase in query, but it did not make better effectiveness.

Keywords: expert search, expert evidence, language modeling method.

1 Introduction

In recent years, much attention has been focused on the task of expert search. A lot of effective models have been proposed, which all made use of the co-occurrence information around expert evidence. This intuition is proved to be quite effective.

The language modeling methods are widely used in expert search. In TREC 2005, Cao et al. [1] and Azzopardi et al. [2] introduced two language modeling methods for the expert search task. These methods were later explained by Balog et al. [3] as candidate model (model 1) and document model (model 2). Fang et al. [4] also proposed a similar framework, but they explicitly modeled on relevance and adopted the probability ranking principle to rank experts.

Further, some detailed problems were studied under the framework proposed in [3]. Petkova et al. [5] proposed a method to consider the dependency between terms and candidates using proximity-based measures. Balog et al. [6] elaborated the estimation of candidate-document association. Serdyukov et al. [7] explored relevance propagation in expert search. Balog et al. [8] considered non-local information available in the collection for expert search. For a thorough review, please refer to [9] and [10].

The language modeling methods have also been proved to be effective under environments other than the TREC collections. For example, Balog et al. [11] created the UvT collection, which involved web pages with multi-linguistic features from a uni-

versity. Besides, Serdyukov et al. [12] and Jiang et al. [13] applied the language modeling method to the internet environment and testified its effectiveness using search engine results.

It is the third year that our group participated in the TREC expert search task. In TREC 2006, we adopted a window-based method for expert search [14]. We firstly built pseudo-profiles for each expert using their co-occurrence information in the documents, and then searched for relevant profiles using text retrieval models. In TREC 2007, we adopted a simple ranking model [15], in which an expert's score is the linear combination of scores for all the supporting documents. In addition, we had adopted several methods to filter out invalid supporting documents, which can effectively enhance precision [16].

This year, we mainly focus on the ambiguity of expert evidence in expert search. In the collection, an expert can appear in several kinds of evidence. Some evidences are ambiguous and can denote more than one expert.

We only consider three main kinds of evidence here:

1. $\boldsymbol{ev_{fn}}$: full name, e.g. "Jiepu Jiang";
2. $\boldsymbol{ev_{abbr}}$: abbreviated name, e.g. "J. Jiang", "Jiang";
3. $\boldsymbol{ev_{em}}$: email address, e.g. "jiepu.jiang@gmail.com".

Generally, ev_{em} is often the most explicit form of evidence, while ev_{fn} and ev_{abbr} can be ambiguous. The ambiguity depends on the size of the collection. For example, ev_{fn} is mostly explicit in an enterprise that involves a few thousand experts, but it can be highly ambiguous over the internet. ev_{abbr} is often highly ambiguous, even only in an enterprise size collection.

Current models [4][5][6][10] consider the ambiguity by involving a combination of several kinds of evidence with different weights in estimating the candidate-document associations. But most experiments indicated that using ev_{abbr} will result in a loss in performance.

In this paper, we focus on better use of the ambiguous expert evidence, i.e. ev_{abbr}. In contrast to other models, our proposed model estimates $p(e|q)$ separately by each kind of evidence that can denote e. Besides, we have also used an auxiliary method in our experiments, i.e. phrase detection in the query.

The remainder of this paper is organized as follows: in section 2, we will describe our methods; in section 3, we will explain our experiments and submitted runs; in the end, we will draw a conclusion and discuss some future challenges.

2 Modeling Expert Evidence in Expert Search

In this section, we will mainly explain our model for expert search, which considers each kind of evidence separately.

From a language modeling perspective, we can rank experts by $p(e|q)$, the probability that the query q is generated by the expert e. Then, applying Bayes rules, $p(e|q)$ can be transformed as Eq. (1):

$$p(e \mid q) = \frac{p(q \mid e) \times p(e)}{p(q)} \tag{1}$$

In a specific ranking task, $p(q)$ is constant for each e and can be ignored in ranking. Besides, we simply assume the same prior probability $p(e)$ for each expert e here. As a result, we can rank experts by $p(q|e)$.

We represent all possible kinds of evidence for e as a set $EV_e\{ev_i\}$, in which ev_i can be any of ev_{fn}, ev_{abbr}, or ev_{em} that can denote e. Further, the whole event space can be partitioned into several subsets for each ev_i. By the total probability formula, we can transform $p(q|e)$ as Eq. (2):

$$p(q \mid e) = \sum_{ev_i \in EV_e} p(q \mid ev_i, e) \times p(ev_i \mid e) \tag{2}$$

In Eq. (2), $p(ev_i|e)$ is the probability that e will appear in the form of ev_i, which can be explained as a kind of expert-evidence association, and $p(q|ev_i,e)$ is the probability that e will generate q when it appears in the form of ev_i, which can be explained as the evidence-topic association.

In the rest of this section, we will further explain our estimatition for $p(ev_i|e)$ and $p(q|ev_i,e)$. For simplification, we have adopted the following assumption here:

Assumption 1: ev_{fn} and ev_{em} are explicit, and only ev_{abbr} can be ambiguous.

In assumption 1, we mean to only consider the ambiguity of ev_{abbr}. The ambiguity of ev_{fn} is ignored here, since it is rare in an enterprise collection that two persons have the same full name.

2.1 Expert-Evidence Association

The association between expert and evidence is measured as $p(ev_i|e)$, the probability that e appears in the form of ev_i. In our model, we estimate it using a maximum likelihood estimation as Eq. (3):

$$p(ev_i \mid e) = \frac{tf(ev_i, e)}{\sum_{ev_j \in EV_e} tf(ev_j, e)} \tag{3}$$

In Eq. (3), $tf(ev_i,e)$ is the frequency of ev_i that denotes e in the whole collection, and Ev_e is the set of all possible evidences that can denote e.

For $tf(ev_{fn},e)$ and $tf(ev_{em},e)$, we can easily count them as $tf(ev_{fn})$ and $tf(ev_{em})$, which are the frequency of ev_{fn} and ev_{em} in the whole collection, since they are assumed to explicitly denote e here (assumption 1).

Now, the rest of the question is how to estimate $tf(ev_{abbr},e)$, which is the frequency of ev_{abbr} that denotes e in the whole collection. The problem is that ev_{abbr} may denote more than one expert and the disambiguation cannot be perfect. For this problem, we first adopt a rule that can disambiguate ev_{abbr} with certainty in a part of the collection, and then use the distribution of ev_{abbr} in the disambiguated part to infer its distribution in the whole collection.

First, though we cannot disambiguate ev_{abbr} in all the documents of the collection, we can disambiguate ev_{abbr} in the documents that contain a related expert's ev_{fn} or ev_{em}. For example, if "Jiepu Jiang" is mentioned in a document, we know that "J. Jiang" in the same document refers to "Jiepu Jiang". Similarly, if "Jay Jiang" is mentioned, we

know that "J. Jiang" in the document is not "Jiepu Jiang". For a simplification, we do not consider the circumstance that both "Jiepu Jiang" and "Jay Jiang" appear in a document. Using this method, we can disambiguate ev_{abbr} in a part of the collection. We represent ev_{abbr} in the disambiguated part of collection as ev'_{abbr}.

Then, we can transform $tf(ev_{abbr},e)$ as Eq. (4):

$$tf(ev_{abbr}, e) = tf(ev_{abbr}) \times p(e \mid ev_{abbr}) \tag{4}$$

In Eq. (4), $tf(ev_{abbr})$ is the frequency of ev_{abbr} in the whole collection (no matter it denotes e or not), and $p(e|ev_{abbr})$ is the proportion of ev_{abbr} that denotes e in the whole collection.

Further, assuming the distribution of ev_{abbr} that denotes e is independent with the disambiguation of ev_{abbr}, we can estimate $p(e|ev_{abbr})$ as Eq. 5.

$$p(e \mid ev_{abbr}) \approx p(e \mid ev'_{abbr}) = \frac{tf(ev'_{abbr}, e)}{tf(ev'_{abbr})} \tag{5}$$

In Eq. (5), ev'_{abbr} refers to ev_{abbr} that can be disambiguated, and $p(e|ev'_{abbr})$ is the proportion of ev_{abbr} that denotes e in the disambiguated part of collection, which can be estimated as the right part of Eq. (5).

2.2 Evidence-Topic Association

The association between evidence and the topic, i.e. $p(q|ev_i,e)$, is the probability that e will generate q when it appears in the form of ev_i. In another perspective, it can also be explained as the probability that ev_i will generate q when it denotes e. We adopt the latter explaination, since it can be easily associated with previous language modeling methods for expert search.

$p(q|ev_{fn},e)$ and $p(q|ev_{em},e)$ can be estimated directly as $p(q|ev_{fn})$ and $p(q|ev_{em})$, since ev_{fn} and ev_{em} are assumed to be explicit (assumption 1). The estimation of $p(q|ev_{fn})$ and $p(q|ev_{em})$ is quite similar to previous models for expert search.

For ev_{abbr}, we can also adopt the intuition in 2.1. First, we can disambiguate ev_{abbr} in a part of the collection. Then, we can estimate $p(q|ev_{abbr},e)$ as Eq. (6):

$$p(q \mid ev_{abbr}, e) \approx p(q \mid ev'_{abbr}, e) = p(q \mid ev'_{abbr-e}) \tag{6}$$

In Eq. (6), $p(q|ev'_{abbr},e)$ refers to the probability of query generation for ev_{abbr} that denotes e in the disambiguated part of collection. For simplification, we use ev'_{abbr-e} to represent ev_{abbr} that denote e in the disambiguated part of collection.

Further, we can estimate the probability of query generation for each of ev_{fn}, ev_{em}, and ev'_{abbr-e} by previous language modeling methods for expert search. In this paper, we only adopted a frequently cited method, i.e. model 2 in [3], as in Eq. (7), Eq. (8) and Eq. (9):

$$p(q \mid ev_i) = \sum_{d_j \in D} p(q \mid d_j, ev_i) \times p(d_j \mid ev_i) \tag{7}$$

$$p(q \mid d, ev_i) \approx p(q \mid d) = \prod_{t \in q} p(t \mid \theta_d)^{n(t,q)} = \prod_{t \in q} \left((1-\lambda) p_{ml}(t \mid d) + \lambda p_c(t)\right)^{n(t,q)} \quad (8)$$

$$p(d \mid ev_i) \propto p(ev_i \mid d) = \frac{tf(ev_i, d)}{\sum_{ev_j} tf(ev_j, d)} \quad (9)$$

In Eq. (8), $n(t,q)$ is the frequency of term t in q, $p_{ml}(t|d)$ is the maximum likelihood estimation for the probability of t in d, $p_c(t)$ is the probability of t in the whole collection, λ is a smoothing parameter which is contantly set to 0.5 here. In Eq. (9), $tf(ev_i,d)$ is the frequency of ev_i in d.

In fact, we can also adopt other language modeling methods for expert search with similar framework to [3], e.g. [5][6][8].

2.3 Other Methods

Some auxiliary methods have been studied in expert search to enhance the effectiveness, e.g automatic detection of expert home page, using HTML structures, the link structure in the collection, etc.

We have also tested for a auxiliary method for expert search. We adopted a method for automatic detecting phrases in the query. For two adjacent terms in the query, i.e. t_i and t_j, we adopt $p(t_it_j|t_i,t_j)$ to determine whether t_it_j is a phrase, which is the probability of t_i and t_j are adjacent when both of them appear in the documents. Further, we adopt a threshold value for $p(t_it_j|t_i,t_j)$ to filter term pairs that are not closely connected. Our previous experiments indicated that this method is profitable in the TREC 2007 collection.

3 Evaluation

We submitted four runs this year to testify our methods proposed in section 2.

First, we submitted a group of runs to testify whether our method considering each expert evidence is profitable:

WHU08BASE: it mostly adopted the model 2 in [3]. But we simplified $p(d|c)$, the candidate-document association. We simply set $p(d|c)$ to 1 if c appears in d, since our previous experiments indicated that this simplification can enhance effectiveness. We used a combination of both ev_{fn} and ev_{em} in this run. ev_{abbr} is ignored since it reduced the effectiveness in our experiments. Phrase detection is used.

WHU08CAN: it adopted the model 2 in [3], and $p(d|c)$ is estimated in Eq. (9). We used a combination of both ev_{fn} and ev_{em} in this run. ev_{abbr} is ignored since it reduced the effectiveness in our experiments. Phrase detection is used.

WHU08RFCAN: it adopted our proposed method to consider each kind of ev_i. We used ev_{fn}, ev_{em}, and ev_{abbr} in run. Phrase detection is used.

Table. 1 gives out an overview of the evaluation results for three runs. We can find out that our proposed method, i.e. WHU08RFCAN, outperformed WHU08CAN.

Table 1. A comparison in previous models and our proposed method.

Runs	MAP	P@5	P@10	R-prec	recip-rank
WHU08BASE	0.3707	0.4255	0.3509	0.3389	0.6563
WHU08CAN	0.3609	0.4509	0.3345	0.3484	0.6296
WHU08RFCAN	0.3765	0.4909	0.3455	0.3579	0.6884

Besides, we have also tested for the phrase detection method:

WHU08NOPHR: it adopted the method in WHU08BASE, but it did not used the phrase detection.

However, our experiments indicated that this method did not achieve better results in TREC 2008, although our previous experiments in TREC 2007 queries showed it is profitable.

Table 2. Evaluation results for the phrase detection method.

Runs	MAP	P@5	P@10	R-prec	recip-rank
WHU08BASE	0.3707	0.4255	0.3509	0.3389	0.6563
WHU08NOPHR	0.3826	0.4909	0.3655	0.3665	0.6770

4 Conclusion

In this paper, we described our methods adopted in the TREC 2008 expert search task.

First, we proposed an adaption to the language modeling method for expert search, which considers the probability of query generation separately using different kinds of expert evidence (full name, abbreviated name, and email). Our experiments indicated that this method can effectively make use of the ambiguous evidence in expert search (e.g. abbreviated name). Previous expert search models can be easily integrated into the method proposed here. Besides, we have adopted a phrase detection method in our experiments, but it failed to make better results.

In the future, we plan to further investigate on the proposed method that considered each kind of evidence. First, we only adopted the most frequently used method for the estimation of $p(q|ev_i)$, which should be tested using more methods. Second, in Eq. (3), Eq. (4), Eq. (5), and Eq. (6), we only used the simple maximum likelihood estimation method, which can be refined in the future.

References

1. Y. Cao, J. Liu, S. Bao, H. Li. Research on Expert Search at Enterprise Track of TREC 2005. In Proceedings of the 14th Text REtrieval Conference (TREC 2005), 2005.
2. L. Azzopardi, K. Balog, M. de Rijke. Language Modeling Approaches for Enterprise Tasks. In Proceedings of the 14th Text REtrieval Conference (TREC 2005), 2005.
3. K. Balog, L. Azzopardi, M. de Rijke. Formal Models for Expert Finding in Enterprise Corpora. In Proceedings of the 29th annual international ACM SIGIR conference on Research

and development in information retrieval (SIGIR' 06), Seattle, Washington, USA, 2006: 43-50.
4. H. Fang, C. Zhai. Probabilistic Models for Expert Finding. In Proceedings of the 29th annual European Conference on Information Retrieval Research (ECIR' 07), Rome, Italy, 2007: 418-430.
5. K. Balog, M. de Rijke. Associating People and Documents. In Proceedings of the 30th annual European Conference on Information Retrieval Research (ECIR' 08), Glasgow, Scotland, 2008: 296-308.
6. D. Petkova, W. B. Croft. Proximity-Based Document Representation for Named Entity Retrieval. In Proceedings of the 16th ACM conference on Conference on information and knowledge management (CIKM' 07), Lisbon, Portugal, 2007: 731-740.
7. P. Serdyukov, H. Rode, D. Hiemstra. Modeling Multistep Relevance Propagation for Expert Finding. In Proceedings of 17th ACM conference on Information and knowledge management (CIKM' 08), Napa Valley, California, USA, 2008: 1133-1142.
8. K. Balog, M. de Rijke. Non-Local Evidence for Expert Finding. In Proceedings of the 17th ACM conference on Conference on information and knowledge management, Napa Valley, California, USA, 2008: 731-740.
9. K. Balog. People Search in the Enterprise. PhD thesis, University of Amsterdam, 2008.
10. K. Balog, L. Azzopardi, M. de Rijke. A Language Modeling Framework for Expert Finding. Information Processing and Management, 45, 2009: 1-19.
11. K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, A. van den Bosch. Broad Expertise Retrieval in Sparse Data Environments. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, 2007: 551-558.
12. P. Serdyukov, D. Hiemstra. Being Omnipresent to be Almighty: The Importance of the Global Web Evidence for Organizational Expert Finding. In SIGIR 2008 workshop on future challenges in expertise retrieval (fCHER), Singapore, 2008: 17-24.
13. J. Jiang, S. Han, W. Lu. Expertise Retrieval Using Search Engine Results. In SIGIR 2008 workshop on future challenges in expertise retrieval (fCHER), Singapore, 2008: 11-16.
14. W. Lu, S. Robertson, A. Macfarlane, H. Zhao. Window-based Enterprise Expert Search. In Proceedings of the 15th Text REtrieval Conference (TREC 2006), 2006.
15. J. Jiang, W. Lu, D. Liu. CSIR at TREC 2007. In Proceedings of the 16th Text REtrieval Conference (TREC 2007), 2007.
16. J. Jiang, W. Lu. IR-Based Expert Finding Using Filtered Collection. In Proceedings of the 4th International Conference on Wireless Communications, Networking and Mobile Computing, Dalian, China, 2008.