IR-Based Expert Finding Using Filtered Collection

Jiepu Jiang, Wei Lu
Center for Studies of Information Resources
Wuhan University
Wuhan, P. R. China
jiepu.jiang@gmail.com, reedwhu@gmail.com

Abstract—**Existing IR-based expert finding generally follows two methods, i.e. the profile-based method and the voting-based one. However, neither the expert-relevant data collected in the profile-based method nor the query-relevant data used for the voting-based method is completely accurate within the confines of current relevance ranking approaches. This problem has been rarely discussed, but impedes expert finding. On this issue, we provide a feasible solution, that is, the collection can be filtered to generate a subset of high-precision relevant data for further processing. In this paper, we propose two perspectives of filtering approaches, i.e. the query-centered perspective and the expert-centered one. For both perspectives, some specific strategies are also discussed and experimented under the CERC collection using the TMJAC model, a voting-based method. On such basis, the different preferences of two perspectives are revealed. Further, to examine the stability of filtering, we examine the filtering strategies using a profile-based method and also testify the effects under the W3C collection. In conclusion, the filtering we proposed is a universal approach of improving expert finding performance.**

Keywords-expert finding; enterprise search; filtering collection

I. INTRODUCTION

Expert finding is an important task of enterprise knowledge management, since the experts constitute an indispensable proportion of the enterprise knowledge. Yimam-Seid et al. identified two different motives in expert finding [1], i.e. listing expertise of the given expert (the information need) and finding experts with the given expertise (the expertise need). The latter motive is the main focus of this paper. For convenience, all the occurrences of the notion *expert finding* in this paper refer to the latter motive specifically.

Early approaches of expert finding are mostly implemented based on structured or semi-structured data source. These applications, though make some achievements, are largely limited by the specific structure of data. On the one hand, it is difficult to completely represent knowledge and expertise in a general format. On the other hand, it is costly to construct such structured data. As a result, expert finding calls for a universal way of processing the heterogeneous information in enterprise.

In recent years, the TREC expert search task began to focus on the IR-based methods of expert finding. Generally, the IR-based expert finding methods can be sorted into two categories, i.e. the profile-based method and the voting-based method. The former method firstly collects expert-relevant data (e.g. terms and document fragments) from the collection to generate expert profiles, and then ranks the experts according to the relevance between their profiles and queries. On the contrary, the latter method firstly retrieves the query-relevant data from the collection, and then uses the retrieved results to vote for experts.

However, both the expert-relevant data collected in the former method and the query-relevant data used for the latter are obtained mainly using IR-based relevance ranking approaches, which inevitably results in irrelevant results and may impede expert finding. On this issue, we are enlightened by the inverse relationship between precision and recall: since the top ranked results of relevance ranking usually receive high-precision, the collection can be filtered by using this feature to generate a high-precision subset. Such subset can better accord with the expert finding methods, but may also perform worse for its lack of data. To fully investigate on this issue, we conduct a research on the methods and effects of filtering in this paper.

The remainder of this paper is organized as follows. In section 2, we have a brief review on the IR-based expert finding. Section 3 explains our intuition and proposes two perspectives of filtering. For both perspectives, some specific strategies are discussed. In section 4, models of expert finding and relevance ranking are explained. Section 5 introduces some details of our experiments. In section 6, results of experiments are evaluated. On such basis, effects of different strategies are compared and the preferences of the two perspectives are discussed. Further, the stability of filtering is also testified. Section 7 draws a conclusion from this research.

II. IR-BASED EXPERT FINDING METHODS

Generally, the IR-based methods of expert finding can be sorted into two categories, i.e. the profile-based method and the voting-based method, which will be explained in this section.

A. The Profile-Based Method

The profile-based method can be described as follows: first, collecting documents or document fragments which are possibly relevant to experts; then, generating a profile for each expert based on the collected data; at last, the experts are ranked according to the relevance between profiles and queries.

The first coming problem of this method is how to collect relevant data for a given expert. A practical consideration is to use the expert home page. Fu et al. performs several ways of detecting home page and proves its effectiveness in expert finding [2]. But it is much more generic to collect information that co-occurs with the expert evidence as the relevant data.

This work was supported by the National Natural Science Foundation of China (No. 70773086).

The collected data can be integrated to generate profiles for experts. Then, the experts can be ranked according to the relevance between profiles and queries. Most intuitively, the profiles can be treated as ordinary documents and the relevance can be estimated by using traditional text retrieval models. But there is no strong theoretical basis to consider expert profiles as ordinary documents. As a result, some improved models have also been proposed to estimate relevance between profiles and queries, e.g. the CDD model [3], which calculates weight for each collected fragment at first and then scores each profile according to the weights of its fragments.

B. The Voting-Based Method

Compared with the profile-based method, the voting-based method follows different intuitions. Basically, it assumes that if a document is relevant not only to a query but also to an expert, the expert and the query are possibly a relevant pair. On such basis, the query-relevant documents can be used as evidence to vote for the experts.

In the voting-based method, the evaluation of relevance between a query q and an expert e generally follows: firstly, associations between e and the documents are calculated; secondly, given q, relevant documents are retrieved with some relevance scores; in the end, the relevance scores of documents which are relevant to e are aggregated as the total relevance score between e and q. This procedure can be formalized as formula 1, where $relevance(e, q)$ stands for the relevance score between e and q, d_i stands for each document in the collection D, $a(d_i, q)$ and $a(d_i, e)$ stand for the strength of association between d_i and q and between d_i and e.

$$relevance(e,q) = \sum_{d_i \in D} a(d_i,q)a(d_i,e) \qquad (1)$$

Specifically, $a(d_i, q)$ and $a(d_i, e)$ can be estimated in various ways. A frequently adopted method is the two-stage language model. Balog et al. [4] have investigated on two different ways of modeling using the language modeling approach, namely the candidate model and the document model. According to their experiments, the latter model performs better than the former and can produce considerable performance.

According to the public results of TREC participants, both methods are effective and prevailing. At present, no adequate evidence reveals that one method prevails against the other.

III. Filtering on Collection

A. Intuition

In section 2, we have a review on two prevailing IR-based expert finding methods. However, both the expert-relevant data collected in the profile-based method and the query-relevant documents retrieved for the voting-based method may fail to be completely accurate due to the IR-based relevance ranking. For the profile-based method, all the documents which contain expert evidence are considered to be relevant to the expert; for the voting-based method, any document that contains query terms is considered to be query-relevant. Apparently, such assumption of relevance is not promising and will inevitably produce irrelevant results.

On this issue, we are enlightened by the inverse relationship between precision and recall: the top ranked results of retrieval can usually produce higher precision. As a result, the collection can be filtered using this feature to generate a high-precision subset, which involves only the top ranked results of relevance ranking. Such high-precision subset can better accord with the expert finding methods and may thus produce a better result.

Considering the problem resides in both query-relevant data and expert-relevant data, the filtering can also be performed in a query-centered perspective or an expert-centered one. For the former perspective, we will rank documents by their relevance to the query and only use the top ranked documents to generate profiles or vote for experts. For the latter perspective, we will rank for each expert his or her relevant documents and only use the top ranked documents.

However, some objections may also exist in the filtering of collection due to its incompleteness. First, for the former perspective, some experts may be excluded from consideration for their absence in any of the top ranked documents, which inevitably reduces recall. Second, only when the expertise information is scattered uniformly in documents of the collection can the subset be equivalent to the full collection in the effectiveness of ranking experts, otherwise the filtering may not work. But hardly can any evidence prove such premise.

To further testify the effectiveness of the filtering intuition, we perform a series of experiments, which are explained in the following sections. The rest of this section discusses some specific strategies for both of the two perspectives.

B. Filtering Strategies

On the issue of this filtering, a practical problem is to find out the appropriate quantity of the top ranked documents that should be used for expert finding. In this section, we will discuss some specific filtering strategies which can be used for both perspectives.

Most intuitively, we can set up a constant cutoff n, which means to directly use the top n ranked results (i.e. the top n strategy). However, for different queries and experts, the retrieved documents may be different, which may make the constant cutoff n unstable. As a result, the *top_percent* strategy is used, where n is a constant proportion of the total quantity of results. Further, the *top_zone* strategy is used, in which, not only the different quantity of retrieved results but also the uncertain distribution of relevance in the retrieved documents are considered. The *top_zone* strategy follows a region analysis way, that is, relevance scores of retrieved documents are accumulated in the ranked order, until the aggregated score reaches a cutoff n, which is a proportion of the total sum of the scores of all the retrieved documents. Besides, sometimes the retrieved results can be very small in quantity, which may make the meaningful proportion only involves one or two documents. As a result, we set a minimum to the quantity of documents, i.e. the *top_percent_min* strategy and the *top_zone_min* strategy.

These specific strategies can be applied in either a query-centered perspective or an expert-centered one. Their effects will be evaluated and discussed in section 6.

IV. MODELS

A. The TMJAC Expert Finding Model

The term voting model with a Jaccard coefficient (TMJAC) for expert finding conforms to the voting-based method and uses a term weighting scheme. In TMJAC, terms are weighted for each expert based on their co-occurrence in each document involved. On such basis, experts are voted by each query term. This model can be formalized as formula 2, where $w(e, t_i)$ stands for the weight of the term t_i for the expert e; w_{ij} stands for BM25 weights of t_i within the document D_j; D_j stands for each document in D', which is a subset of the collection; $J(e, t_i)$ is the Jaccard coefficient of the co-occurrences between e and t_i.

$$w(e,t_i) = \sum_{D_j \in D'} w_{ij} \times J(e,t_i) \tag{2}$$

Note that D' refers to a conditional subset of collection in which all the documents are involved to vote for $w(e, t_i)$. In a baseline model without filtering, D' refers to a subset in which all documents contain both e and t_i. But the filtering strategy will conditionally change D' for a better results.

BM25 [5] is a popular probabilistic model of information retrieval, in which terms are weighted for each document. The BM25 weight w_{ij} of term t_i in a document D_j can be formalized as formula 3, where tf_{ij} stands for the raw term frequency for t_i in D_j, dl_j is the length D_j, $avdl$ is the average length of documents in collection, n is the document frequency for t_i, N is the total document count in the collection, k_1 and b are two parameters, which are set to 1.2 and 0.75.

$$w_{ij} = \frac{(k_1+1)tf_{ij}}{k_1(1-b+b*\frac{dl_j}{avdl})+tf_{ij}} \log\frac{N-n+0.5}{n+0.5} \tag{3}$$

The Jaccard coefficient is used for evaluating the similarity between two sets. Specifically, in the TMJAC, the Jaccard coefficient $J(e, t_i)$ between an expert e and a term t_i stands for the strength of co-occurrences for e and t_i, see formula 4.

$$J(e,t_i) = \frac{\left|D_e \cap D_{t_i}\right|}{\left|D_e \cup D_{t_i}\right|} \tag{4}$$

In formula 4, $|D|$ stands for the number of document inside the collection D; D_e stands for the subset which contains evidence of e; D_{ti} stands for the subset which contains occurrence of t_i. Table 1 shows a comparison of effectiveness between TMJAC and a term voting model without considering the Jaccard coefficient (TM), which reveals that the Jaccard coefficient can enhance the effectiveness in a term weighting model under the CERC collection.

TABLE I. COMPARISON BETWEEN TMJAC AND TM

Model	MAP	R-prec	Bpref	recip-rank	P@5	P@10
TMJAC	0.2384	0.2032	0.6039	0.3215	0.120	0.080
TM	0.2010	0.1589	0.6172	0.2745	0.120	0.076

B. Relevance Ranking Models

For the query-centered perspective of filtering, we will rank the documents by their relevance to the query. The relevance ranking model used to retrieve the query-relevant documents is BM25, which has been explained in the previous subsection.

For the profile-based method, we will rank for each expert his or her relevant documents. The relevance between experts and documents can be estimated by means of considering the expert evidence as query terms. If all variations of the expert evidence are treated the same, it can be formalized as formula 5, which is transformed from the *tf* module in BM25 to normalize the term frequency inside the documents. In formula 5, $w(e_i)$ stands for relevance between the expert e and the document D_i, f_i is the raw frequency of e, dl_i is the length of D_i, k_1 and b are two parameters, which are set to 1.2 and 0.75.

$$w(e_i) = \frac{(k_1+1)f_i}{k_1(1-b+b*\frac{dl_i}{avdl})+f_i} \tag{5}$$

V. EXPERIMENTS

Our experiments are mainly based on Lucene under a Windows environment. However, the default mechanism for indexing and ranking in Lucene only conforms to the Vector Space Models. To support indexing and retrieval for other IR models, we developed Lucene-Ex [6], an extensive plug-in for Lucene, which provides indexing and ranking function for BM25 and Language Model.

The collection to testify the effects of filtering is the CERC collection [7], which is used in TREC 2007 enterprise track. To examine the stability of filtering, the W3C collection is testified, which is used in TREC 2005 and 2006.

The expert recognition process for each collection is implemented in a rule-based Named Entity Recognition method, which is similar to Mikheev et al. [8]. But only the expert full name and email address are considered as expert evidence.

Experiments involved in section 6.1 and section 6.2 use the TMJAC model under the CERC collection. In section 6.4, a profile-based method and the W3C collection will also be used to examine the stability of filtering.

VI. EVALUATION

A. Query-Centered Filtering

For the query-centered perspective, the strategies proposed in section 3.2 are testified by using TMJAC under the CERC collection. To make a comparison, the baseline run is shown in the following experiments, which are the evaluation results provided in section 4.1 using TMJAC.

Fig. 1 shows the MAP results of the top n strategy with the cutoff n ranging from 1 to 100. It is revealed that the top n strategy is distinctly fruitful. Note that even when the cutoff n is set to 1, which means only voting for the experts contained in the first document retrieved for each topic, the filtering run receives a better result than the baseline run.

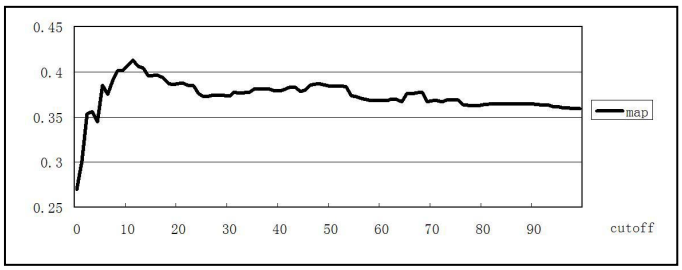

Figure 1. MAP of filtering using the top *n* strategy.

Basically, the effectiveness of filtering has been testified, but the constant value of cutoff *n* is rough. Fig. 2 gives the MAP result using the *top_percent* strategy and the *top_zone* strategy, where the proportion ranges from 1/500 to 1/20. Note that the proportion is represented as a fraction, which results in an asymmetry x-axis. It is shown in the experiments that the *top_zone* strategy performs better than the *top_percent* strategy, which reveals that it is reasonable to consider the uncertain distribution of relevance score among retrieved results.

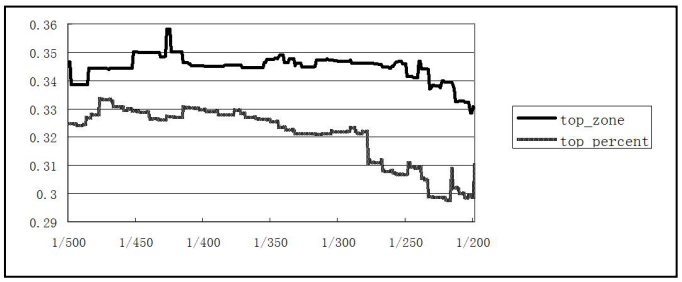

Figure 2. MAP of filtering using *top_zone* strategy and *top_percent* strategy.

However, compared with the top *n* strategy, the maximum MAP received by the *top_zone* strategy is decreased. In order to make clear the problem, we have a further investigation. The 50 queries in the CERC collection are divided into two groups. Group A contains 34 queries with more than 5000 retrieved documents, while Group B contains 11 queries with less than 1000 retrieved documents. The rest of the queries are excluded. Note that the average retrieved document number for Group A and Group B is 33947 and 341.

In this way, we discover that the very cutoff proportion which produces the best performance for Group A results in a relatively worse performance in Group B, because queries in Group B at the same proportion contain merely one or two documents. This phenomenon proves our concerns for the scarcity of retrieved results in section 3.2.

As a result, the *top_zone_min* strategy is experimented. Fig. 3 gives the MAP results using the *top_zone_min* strategy setting different minimums of 5, 10, 15 and 20, in which the zone proportion ranges from 1/500 to 1/20. It is revealed that the combination of the top *n* strategy into the *top_zone* strategy is advisable for improving effectiveness of the *top_zone* strategy.

Though still no evidence shows that using a variable cutoff size will overwhelm a simple cutoff of constant *n*, theoretically, the *top_zone_min* strategy may be more stable than the top *n* strategy, since it concerns both the different quantity and the variable distribution of relevance in retrieved results.

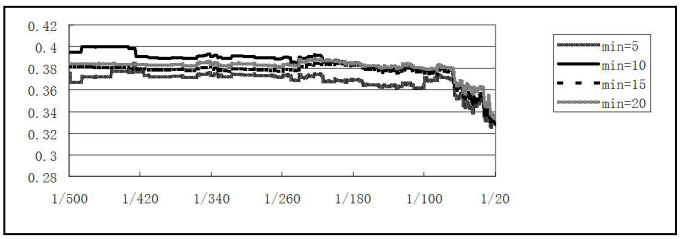

Figure 3. MAP of filtering using the *top_zone_min* strategy.

B. *Expert-Centered Filtering*

The filtering strategies proposed in section 3.2 are proved in a query-centered perspective. In this section, we will further testify the strategies in an expert-centered perspective.

Considering that for most of the experts the retrieved documents contain less than 20 results, the variable size of cutoff is dispensable. Another consideration is whether to restrict that the retrieved documents for each expert should also contain query terms or not, i.e. the *exp_tp_top_n* strategy. Fig. 4 shows MAP results of two filtering strategies under the expert-centered perspective. It is showed that both strategies are fruitful. In comparison, the *exp_top_n* strategy can receive better results than the *exp_tp_top_n* strategy.

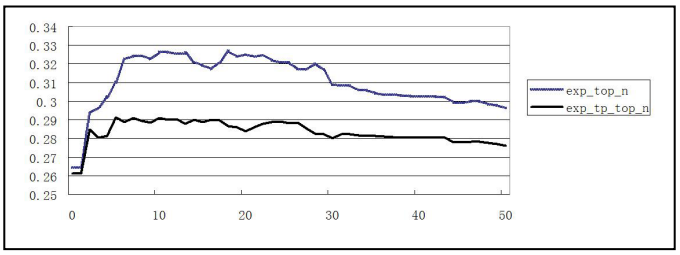

Figure 4. MAP of filtering with an expert-centered perspective.

C. *Comparisons between Two Perspectives*

The effectiveness of both the query-centered filtering and the expert-centered filtering has been proved by experiments. However, another phenomenon arouses our attention, that is, the quantities of returned experts in these two perspectives of filtering generally have different trends. As what is revealed in Fig. 5, for the top *n* strategy, the number of relevant experts returned (i.e. *rel-ret*) at peak MAP value is relatively smaller. But with an increasing number of documents taken into account, though *rel-ret* increases, MAP goes down. However, for the expert-centered filtering, *rel-ret* has the same trends with MAP.

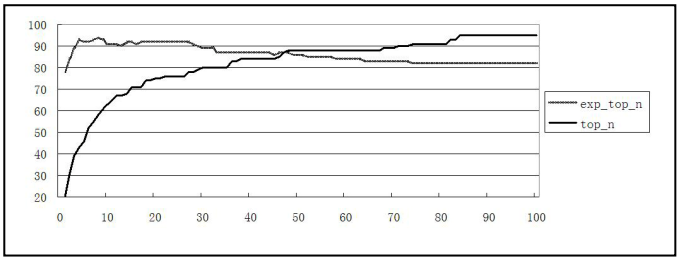

Figure 5. Different trends of *rel-ret* for two perspectives of filtering.

A relatively smaller *rel-ret* but a higher MAP shows that the top *n* strategy can produce results more precise in query-centered perspective than in expert-centered perspective. However, though the maximum of MAP received by the *exp_top_n* strategy is relatively lower, it returns a larger list of relevant experts. To conclude, the filtering on collection has different preferences in different perspectives, that is, the query-centered perspective prefers precision while the expert-centered one can promote recall. Our experiments also showed that the query-centered perspective of filtering is more effective in enhancing MAP than the expert-centered perspective.

D. Stability of Filtering

We have already testified the effectiveness of the filtering on both perspectives using TMJAC model under the CERC collection. But still more works are required to prove the effects of filtering to be universal. Firstly, other expert finding methods should be used. Secondly, more collections should be examined.

For the first part, we adopt a profile-based method to examine the effect of the top *n* strategy. The profile-based method which we adopt uses a window size of 100 words to collect information around the expert evidence. The top *n* strategy is used to select the subset in a query-centered perspective. Then, the collected data is combined literally and BM25 is used to retrieve the profile. Fig. 6 shows the MAP results of the top *n* strategy using the profile-based model. It is revealed that the top *n* strategy also gives a distinct improvement to the profile-based method.

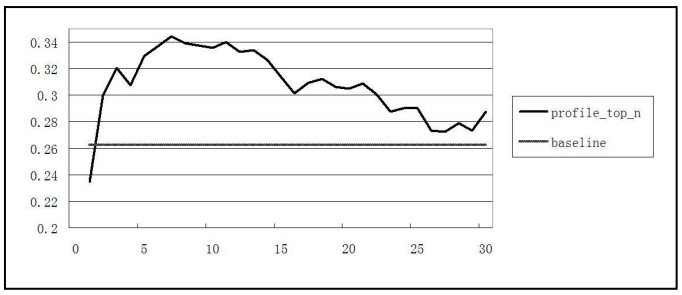

Figure 6. MAP of the top *n* filtering strategy using a profile-based method.

Further, we have also tested for the W3C collection based on TMJAC model and the top *n* strategy, which is revealed in Fig. 7. It is shown that the filtering is also distinctly effective under the collection other than CERC.

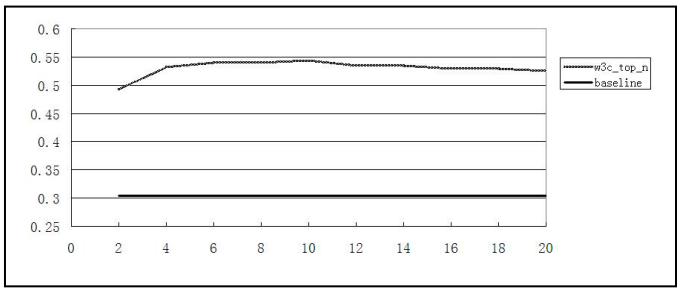

Figure 7. MAP of the top *n* filtering using TMJAC under W3C collection.

However, we have also found that both perspectives of filtering cannot produce distinct effects using the two-stage language model [4], though the different preferences still exist. A possible explanation to this problem is that the relevance score in the two-stage language model varies a lot in quantity, usually ranges from 10^{-4} to 10^{-16}, which makes the lower ranked results have in fact little impact on the vote for experts. As a result, it makes little difference to filter or not.

To conclude, the stability of filtering is generally examined. It is revealed that the filtering is effective whenever using a profile-based method or a voting-based method, though not for the two-stage language model. The effects of filtering are also stable under collections other than the CERC collection.

VII. Conclusion

In this paper, we have an investigation on the effects of collection filtering for expert finding. Two perspectives of filtering are proposed, i.e. the query-centered perspective and the expert-centered perspective. For both perspectives, a few specific strategies are discussed and proved to be effective using the TMJAC model under the CERC collection. By comparison, we also discuss that the former perspective prefers to precision while the latter promotes recall. Further experiments have been made to examine the stability of filtering. It is testified that the filtering strategies are effective also in a profile-based method and under the W3C collection. But future research is needed to better explain why the filtering strategies do not work well using the two-stage language model. However, the filtering we proposed is generally testified to be effective in different expert finding methods and collections.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 70773086).

References

[1] D. Yimam-Seid and A. Kobsa. Expert finding systems for organizations: problem and domain analysis and the DEMOIR approach. Journal of Organizational Computing and Electronic Commerce, 13(1):1-24, 2003.

[2] Y. Fu, Y. Xue, T. Zhu, Y. Liu, M. Zhang and S. Ma. THUIR at TREC 2007: enterprise track. In Proceedings of the sixteenth Text REtrieval Conference (TREC 2007), Gaithersburg, USA, 2007.

[3] Y. Fu, R. Xiang, Y. Liu, M. Zhang and S. Ma. A CDD-based formal model for expert finding. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal, 2007, pp.881-884.

[4] K. Balog, L. Azzopardi and M. 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, Seattle, USA, 2006, pp.43-50.

[5] S. Robertson, H. Zaragoza and M. Taylor. Simple BM-25 extension to multiple weighted fields. In Proceedings of the thirteenth ACM international conference on Information and knowledge management, Washington, D.C., USA, 2004, pp.42-49.

[6] http://sim.whu.edu.cn/newsim/lucene-ex/index.html

[7] P. Bailey, N. Craswell, I. Soboroff and A. Vries. The CSIRO enterprise search test collection. ACM SIGIR Forum, 41(2), 2007, pp.42-45.

[8] A. Mikheev, M. Moens and C. Crover. Named entity recognition without gazetteers. In Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics, Bergen, Norway, 1999, pp.1-8.