Enhanced Algorithms for Enterprise Expert Search System

Valentin Molokanov, Dmitry Romanov and Valentin Tsibulsky
National Research University "Higher School of Economics",
Moscow, Russia

ABSTRACT

We present the results of our enterprise expert search system application to the task introduced at the Text Retrieval Conference (TREC) in 2007. The expert search system is based on analysis of content and communications topology in an enterprise information space. An optimal set of weighting coefficients for three query-candidate associating algorithms is selected for achieving the best search efficiency on the search collection. The obtained performance proved to be better than at most TREC participants. The hypothesis of additional efficiency improvement by means of query classification is proposed.

Keywords: TREC, expert search, enterprise information management

1. INTRODUCTION

In large organizations information repositories are characterized by an extreme heterogeneity of structured as well as unstructured information in a vast amount of variously formatted documents. Indeed, different reports, meeting protocols, descriptions of working groups, projects, etc. are valuable sources which can be found via some usual search functionality. Despite of that, a user has to spend too much time in viewing such documents in order to find employees who are competent at a given theme at the present time. The task of finding people with concrete professional experience arises unavoidably in the need of asking anything in some professional area as well as in performing a series of other more difficult tasks; among them are, for example, finding all members of a specified project or finding all employees that are working with a specified customer. In similar scenarios using an enterprise expert search system is more advantageous in comparison with a simple search engine, as the user can find the appropriate people much faster. An expert search system delivers a list of people who might have knowledge and be useful as experts at a given topic. So an expert search system can be an effective means of organization management in the purposes of improving performance and collaboration quality by presenting information about the employees who possess knowledge in requested areas.

We continue our research of expert ranking algorithms using the TREC Enterprise track 2007 corpus. The corpus represents the crawl of the open-access information from the official site of the Commonwealth Scientific and Industrial Research Organization (CSIRO) [1]. Although in 2009 TREC Enterprise track was replaced by Entity track to change a research direction towards finding arbitrary entities in web data, the Enterprise track collections remained accessible, so we decided to use the 2007 corpus to optimize our expert search system performance.

The main problem for any automated expert search system is to associate a query with people. To identify such associations, various techniques are proposed. Expert search methods in modern enterprise systems are rather different, so there is no conventional expert search model for enterprise systems. However, most of them could be classified into two principally different approach types: document-based and candidate-based.

The document-based approach became the first acceptable approach to expert search. It imitates expert search process with the use of an ordinary search system. Here, the primary retrieval of relevant documents and the following people search in such documents are implied. The approach is just referred to as a two-stage model and is described in details in [2]. It became widely used in its several variants at TREC 2007 [3]. Another quite natural approach to expert search is a candidate-based one. It supposes building a special description (so-called profile) for each candidate, after that candidate ranking is produced with the help of simple search technologies. With the help of various methods candidates' profiles are filled with their expertise information. The examples of such methods are presented in [4], [5], [6], [7]. Despite the variety of interpretations of query-candidate connections in these two approaches, TREC 2008 results showed that the mentioned approaches have no advantages over each other, and the best expert search efficiency at TREC was achieved by regarding structured information in documents [8] and intranet structure [9], as well as by handling additional information from beyond the collection [10].

International Conference on Graphic and Image Processing (ICGIP 2012),
edited by Zeng Zhu, Proc. of SPIE Vol. 8768, 876836 · © 2013 SPIE
CCC code: 0277-786X/13/$18 · doi: 10.1117/12.2010936

We decided to follow another way. Our model's idea consists in the possibility that expert search process can be organized without preliminary finding documents on the requested topic. At the same time, the model does not need to address any structured text fragments or external data. Although in this sense our model is simpler than candidate-based models demonstrated at TREC, it does not lose at efficiency to them. A high expert search efficiency is reached in our model due to several enhanced algorithms simulating a query-candidate association. It can be said that we propose an alternative approach to expert search as compared with TREC participants' models. A brief description of our model and the results obtained on the TREC 2007 expert search task are given below.

2. EXPERT SEARCH MODEL

Our model is essentially candidate-based. Indeed, it saves information about terms and their positions in documents, however the model is attached to the set of terms the candidate "said" in the collection, rather than to the documents. This is a unique model's feature, so our model is sharply different from expert search models demonstrated at TREC.

To improve expert search efficiency we apply the several specific techniques. These are the following.

1) Term weighing. For each term in the collection we assign its significance. The significance of the term is its natural weight feature that is connected with its statictical properties in the collection. The employment of significance allows us to effectively distinguish a professional lexicon from a common-used one.

2) Building associative connections of a candidate with terms and bigrams. In order to calculate a term-candidate (or bigram-candidate) association measure we take into account the frequency of term usage by a candidate, the amount of sent and received messages containing the term as well as the amount of people with whom a candidate exchanges such messages. All this we performed earlier [11], and now, besides, we also recognize email addresses in the text and associate the terms standing near an address to the person who has this address according to the collection mapping. This is a modification of our model as compared with its previous version.

3) Building associative connections between terms. We introduce a term-term connection cardinality and define it based on how close to each other these terms appear in the original texts. For each significant term we construct the set of expanding terms, i.e. terms which are connected with this term. As a result, a query can be automatically expanded by mentioned terms: a user may get proper experts even by specifying an implicitly close query, he does not need to specially select the terms characterizing those experts.

4) Combining several expert ranking ways. We use expert ranking based on three algorithms that identify people connections with terms, expanding terms and bigrams respectively. So we calculate the values of three expert rating parameters. And the resulting expert rank is defined as a linear combination of these three parameters, with three corresponding weighting coefficients being specified as system settings. Thus, by using three weighting coefficients we merge three expert ranking algorithms into a single weighting expert search model.

A detailed description of our model requires a scope of a special paper which we are going to publish in nearest time.

3. RESULTS

We compared and optimized search results with the help of a specially prepared high-performance user application. It enables to conduct multiple runs of our system with various sets of parameters. From run to run we changed weighting coefficients for considered lexical types of ranking (query terms — C_t, expanding terms — C_e, bigrams — C_b), and also varied number of the expanding terms involved in calculations.

We found those sets of setting coefficients at which the best mean average precision (MAP) values are reached. Table 1 lists such optimal sets of settings as well as the corresponding expert ranking scores for each run (here the parameter *l* is expanding terms cutting level by their significance, and P@5 and P@20 are search precisions at 5th and 20th ranks, respectively). The first two runs, HSE2007q and HSE2007qn, relate to our previous model that does not associate terms with people on basis of their email addresses. It is worth mentioning that we performed our runs on both short queries (q) and queries with narratives (qn). We optimized our modified model on the same query types and presented the results of this optimization in the last two rows of Table 1. We see that MAP on short queries has improved and MAP on narrated queries has deteriorated, but both changes are slight. The detailed analysis of answers revealed these MAP changes occurring due to changes of answer precision on few queries: for example, in the short query run the answer precision is increased greatly for the 49th query "atmosphere", and for other queries its highest changes are only

in the second — third digit. It is also important that our model's modification exerts the greatest influence on single-term queries. One could quite consistently explain this fact. Long (both multiple-term and narrated) queries contain a sufficient number of the significant terms defining the subject matter of search, while a single word can be insufficiently significant for this. So recognition of an email address as a term (for modeling term-candidate associations via this email) seems to be reasonable when a query is short.

Table 1. Optimal weighting coefficient values and expert ranking scores

Run	C_t	C_e	C_b	l	MAP	P@5	P@20
HSE2007q	5	0.1	10	5	0.3655	0.192	0.079
HSE2007qn	0.0001	1	0.5	100	0.3622	0.188	0.078
HSE2007qMod	4.7	0.1	8.5	5	0.3771	0.200	0.083
HSE2007qnMod	0.1	10	10	100	0.3569	0.188	0.077

We can compare the amount of data stored in the system before and after the performed modification. The revision of the term-candidate associating method and the corresponding index rebuilding increased the total number of terms in our system by 14%. This doubled the number of term-term, term-person and bigram-person interconnections (Table 2). Thus, an almost negligible precision increase due to a relatively small change of the number of terms gives rise to an enormous increase of the required data space. Although our system did not lose in its performance in this time, the further system overflow will finally have a definite impact on both the system performance and the server free memory.

Table 2. Index statistics for two expert search models

Data object	Initial model	Modified model
terms	500 888	571 147
communications	325 723	325 723
bigrams	4 699 190	9 539 334
term-term associations	73 312 432	157 390 218
term-person associations	21 821 770	44 530 076
bigram-person associations	103 759 630	209 956 386

Notice that there are some inaccuracies in the TREC 2007 relevance judgements file mapping. Some email addresses which are mapped in the relevance judgements file as relevant experts were found to be non-existing in the CSIRO collection. We registered several cases of email misprinting, usage of different name forms, as well as the belonging one email address to several people. Emphasize that the MAP values listed above are actually underestimated owing to the fact that we take these non-existing experts into account in MAP calculations. Despite of understated precision evaluations, the results of our runs confidently get into the top half of the TREC 2007 results table (see Table 4 in [3]).

We continue our research connected with the possibility of expert search efficiency improvement in terms of query assessment and classification. We suppose that some characteristic of query quality can be formed based on internal query properties. Here we demonstrate it on the example of single-term queries. If these queries are ordered by their term significance then a distinct idea about the necessity to narrate a query becomes apparent (Fig. 1, for the sake of scale conservation a response parameter proportional to term significance is presented instead of significance itself). For low-significant query terms the narrative proves to increase the answer precision (the answer precision difference bars are directed upwards). This implies that if query terms are common-used, the answer precision on this query is low and the query is to be narrated; one can expect the narrative to this query to contain more significant terms. On the contrary, in the right part of Fig. 1 the high-significant queries are presented. There is no need to narrate them because their terms are specialized enough (relative to the collection) for the necessary thematic field to be determined and a quite accurate expert selection to be provided. Thus, in two outermost significance zones we can unambiguously say whether there is a necessity to narrate a query. Notice that in this approach we are basing only on internal parameters calculated in the system without making reference to the relevance judgements file.

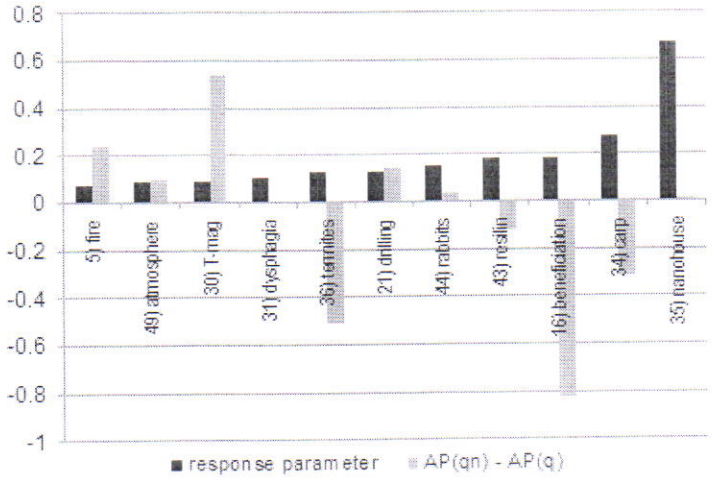

Figure 1. Response parameter on single-term queries and precision change due to narrating

Using a query quality it is principally possible to create some preliminary automatic query classification mechanism. For example, using some assessments the system could conjecture that the query refers to such a query category in which answers usually do not give high precision; thereafter the system could suggest the user to add a narrative as an explanation for specifying the query. The further performance during the expert search could suppose narrative assessments, on the base of which the most proper ranking variant could be provided to maximize the answer precision. Such a two-stage algorithm can highly increase the expert search precision in rather general topics or with a large number of relevant experts in a particular sphere.

4. CONCLUSIONS AND FUTURE EXPLORATIONS

We performed the second stage of optimization of our enterprise expert search system with the use of TREC data. Remaining in the scope of our approach to use no data from beyond the collection, we found some internal reserves in the data and used them in order to establish a more accurate conformity between people and query terms. This enabled to increase the search precision by 3%.

We have further perspectives for optimization of search quality in our enterprise expert search system. From one hand, we can continue to accumulate data in the system in the form of terms. The procedure which was performed by many TREC participants gave no appreciable precision growth, but it doubled the data space necessary to store in the system. Really, an increase of the number of terms in the system results in a much more increase of the amount of bigrams and, especially, the amount of term-to-term connections. If we want to continue optimization in this direction, we will have to modify our model in order to avoid index overflow and performance loss.

Besides, we plan to optimize our system towards settings individualization for each query. We showed earlier that about a half of all CSIRO queries can be implemented at constant settings, and other CSIRO queries should be handled at individual settings, i.e., with domination of one or two of the three considered expert ranking algorithms. We suppose that the choice of the most appropriate algorithm must depend on query quality. The issue about query quality requires further exploration. What is the criterion of a "good" query formulation, how complete must user's information be for asking the system, how can an effective query modification suggestion be formed based on system response — this is to be clarified during more detailed exploration of interaction between our system and a mapped text corpus.

5. ACKNOWLEDGEMENT

This work was conducted with financial support from the Government of the Russian Federation (Russian Ministry of Science and Education) under contract 13.G25.31.0096 on "Creating high-tech production of cross-platform systems

for processing unstructured information based on open source software to improve management innovation in companies in modern Russia".

REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff and A.P. de Vries, "The CSIRO Enterprise Search Test Collection", SIGIR Forum **41**, 42-45, (2007).

[2] Y. Cao, J. Liu, S. Bao, H. Li and N. Craswell, "A Two-Stage Model for Expert Search", Tech. Report MSR-TR-2008-143, Microsoft Research, (2008).

[3] P. Bailey, N. Craswell, A.P. de Vries and I. Soboroff, "Overview of the TREC 2007 Enterprise Track", Proc. TREC-16, 30-36, (2007).

[4] Z. Ru, Q. Li, W. Xu and J. Guo, "BUPT at TREC 2006: enterprise track", Proc. TREC-15, 151-156, (2006).

[5] W. Lu, S. Robertson, A. Macfarlane and H. Zhao, "Window-based enterprise expert search", Proc. TREC-15, 186-193, (2006).

[6] Y. Fu, W. Yu, Y. Li, Y. Liu, M. Zhang and S. Ma, "THUIR at TREC 2005: enterprise track", Proc. TREC-14, 772-779, (2005).

[7] G. You, Y. Lu, G. Li and Y. Yin, "Ricoh research at TREC 2006 enterprise track", Proc. TREC-15, 570-582, (2006).

[8] K. Balog and M. de Rijke, "Combining Candidate and Document Models for Expert Search", Proc. TREC-17, 328-331, (2008).

[9] J. Yao, J. Xu and J. Niu, "Using role determination and expert mining in the enterprise environment", Proc. TREC-17, 173-178, (2008).

[10] K. Balog, I. Soboroff, P. Thomas, P. Bailey, N. Craswell and A.P. de Vries, "Overview of the TREC 2008 enterprise track", Proc. TREC-17, 14-25, (2008).

[11] V. Molokanov, D. Romanov and V. Tsibulsky, "Optimization of Algorithms and Parameter Settings for an Enterprise Expert Search System", Proc. ICCSISCT, to be published.