Optimization of Algorithms and Parameter Settings for an Enterprise Expert Search System

Valentin Molokanov, Dmitry Romanov, Valentin Tsibulsky
National Research University "Higher School of Economics"
Moscow, Russia
vmolokanov@it.ru, dromanov@hse.ru, vtsibulsky@hse.ru

Abstract *:*

We present the results of our enterprise expert search system application to the tasks that were introduced at the Text Retrieval Conference (TREC) in 2005—2007. The expert search system is based on the analysis of content and communications topology in an enterprise information space. During the performed experiments an optimal set of weighting coefficients for three query-candidate associating algorithms is selected for achieving the best search efficiency on a specified corpus. 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*

I. INTRODUCTION

Finding people with concrete professional experience is one of the most actual tasks in the field of enterprise content management. It 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 response with enumeration 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.

The expert search task state is universal and simple: the system must find potential candidates and arrange them in descending order of their theme expertise probability (in other words, rank them) using the corpus data.

In 2005 expert search became one of the official tasks in the TREC Enterprise track. This research area provided a general experimental base consisting of the following main elements: the collection of documents, the topic list and the list of experts in each topic (so-called relevance judgments file). The first TREC Enterprise track collection includes the public documents of the World Wide Web Consortium (W3C) [1], most of which represent email messages, and in 2007 a new corpus was introduced into these experiments—this is the Commonwealth Scientific and Industrial Research Organization (CSIRO) enterprise collection [2] which is the crawl of the open-access information from the CSIRO official site.

As for query-candidate associations identification, TREC participants proposed various techniques. Nevertheless, overwhelming majority of them realize two principally different approach types: document-based

and candidate-based. In the first case the primary retrieval of relevant documents and the following people search in such documents are implied, that is, the document-based approach imitates expert search process with the use of an ordinary search system. The candidate-based approach supposes building a special description (so-called profile) for each candidate, after that candidate ranking is produced with the help of simple search technologies.

In our expert search model we accept a candidate-based idea and, in addition, propose some innovations designed for expert search efficiency improvement. Our model's novelty consists in using the following techniques.

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. As a term-candidate (or bigram-candidate) association measure, we introduce a connection cardinality (i.e., power) between them. The frequency of term usage by a candidate, the amount of sent and received messages containing this term as well as the amount of people with whom a candidate exchanges such messages — all listed contributes to the term-candidate connection cardinality.

3) Building associative connections between terms. We introduce a term-to-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.

We apply the proposed model to the TREC Enterprise track expert search task 2005—2007. In reproducing each of the tasks we empirically select the optimal combinations of model parameters for reaching the best expert search efficiency. The experimental results are evidence of a reliable expert search quality reached by our system and, in addition, reveal the potential for its further improvement.

II. RELATED WORKS

For solving an expert search task with the help of automated systems many expert search engine models were developed by different TREC participants. The basic model which was carried to an acceptable completion level enough satisfactory in 2005 is referred to as a two-stage expert search model [3]. Many TREC participants used some variant of this model [4] along with their own supplements integrated with it. The two-

stage expert search model is document-based and implies two stages in obtaining the resulting expert list: these are document search (referred to as relevance stage) and people search in documents (co-occurrence stage).

At the document search stage the document relevance relative to the query is evaluated. It is clear that such functionality was realized much earlier in open-access search systems, and most of document search algorithms here come to calculating the ratio of term usage frequencies in the document and in the collection. Really, if the word is used in the document more often, there arises a natural reason to consider that this document is more concerned to a requested topic.

The people search stage is of an innovative interest. Here a query-candidate association extent is evaluated. The higher is this extent the more relevant a candidate is thought to be. Emphasize that there is no single generally accepted algorithm for expert search, so a query-candidate association is regarded in several ways. One of the most popular approaches consists in the following: the smaller is a text window containing both query terms and candidate mention, the higher is a query-candidate connection level. This is known as a window-based model (or proximity model) which is used by some TREC groups [5], [6], [7] as the main technique for candidates' expertise assessment. From other well-established expertise evaluation approaches, document mapping oriented methods are also often-used. Generally, these methods are based on HTML mapping for web pages, some special fields for email messages, etc. A particular example is a title-author model [3]. Here, a candidate is considered to be relevant to the topic if he is an author of a document and the query terms appear in the title of this document.

Some smaller amount of TREC participants held the candidate-based approach. With the help of various methods candidates' profiles are filled with their expertise information. Expert search techniques listed above are also applied in a candidate profile building process. For example, in the early versions of TREC expert search task (2005—2006) the prevailing techniques were term-candidate proximity-based model [8], [9] as well as using structured information from web pages, such as document title, headers at various levels, other bold-facing text strings, etc. [8], [10], [11]. Later, in the 2008 expert search task, there appeared a greater variety of techniques [12]. The were demonstrated expert identifying methods aimed at treatment of different types of candidate mentions, link analysis, knowledge areas extraction from beyond the collection, finding candidates' homepages and intranet structure consideration, combining document-based and candidate-based models.

To summarize, we should say that expert search methods in modern enterprise systems are rather different, so there is no conventional expert search approaches for enterprise systems. We have developed our own enterprise expert search system and apply it to the TREC Enterprise track expert search task 2005—2007. A brief description of our model and the corresponding experimental results are given below.

III. EXPERT SEARCH MODEL

Our model's idea consists in the possibility that expert search process can be organized without preliminary finding documents on the requested topic. Our model is essentially candidate-based. Indeed, we save information about terms and their positions in documents, however the model becomes attached to the set of terms the candidate "said" in the collection, rather than to the documents. This is a unique model feature, so our model is sharply different from expert search models demonstrated at TREC.

We use a specific form of the proximity-based model for calculating term-to-term connections. The approach based on two arbitrary semantic constructions proximity in the text is general enough: using it one can associate these constructions even without taking structure of documents, paragraphs, sentences into account. Therefore it may be adapted for solving a lot of problems arising during unstructured information processing. In our expert search system we apply the proximity-based model for associating terms, whereas among many other TREC participants it was used to identify term-candidate associations. Besides, the proximity model turned out to be quite successful in such challenges as fact-based information extraction, entity categorization, clusterization and selecting keywords for describing relations between similar entities [13].

Term-candidate associations are modeled by means of analyzing lexical composition of a collection and calculating term usage statistics for people. Between each significant term (or bigram) and each person we calculate a corresponding connection cardinality. Moreover, we perform similar calculations not only for query terms, but also for terms appearing nearby them in documents (so-called expanding terms) and for bigrams appearing in a query, thereby our model combines three corresponding expert ranking algorithms for which a user is able to assign their weighting coefficients as system settings. Thus, we perform expert ranking as a linear combination of three lexical parameters calculated in our system.

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

IV. RESULTS

To compare and optimize search results, multiple expert search system runs with various sets of parameters were carried out in a specially prepared high-performance user application, and for each run the values of search precision metrics accepted on TREC were fixed. These are mean average precision (MAP), precision at 5th (P@5) and 20th (P@20) ranks [14]. From run to run we changed weighting coefficients for considered lexical types of ranking (query terms, expanding terms, bigrams), and also varied number of the expanding terms involved in calculations.

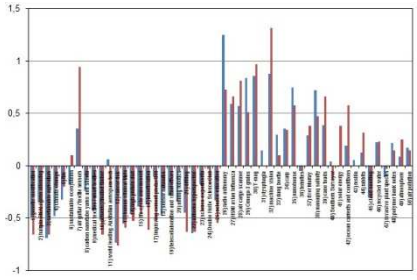
In performing automatic runs we found some universal sets of settings for all queries of a collection. Table I lists optimal weighting coefficient values for three lexical ranking types (C_t – query terms, C_e – expanding terms, C_b – bigrams, l – parameter of expanding terms cutting by their significance level) in each automatic run (q – short query, qn – query with narrative) and corresponding precision factors for runs. The presented settings give best MAP value in comparison with other possible settings options; in addition, other considered precision indicators also appear near an optimum. Comparing results of our runs on the 2005, 2006 and 2007 corpora on short queries (q) to TREC participants' results (see, respectively, Table 5 in [1], Table 4 in [15] and Table 4 in [4]), it is possible to conclude that the shown experts search accuracy surpasses the accuracy obtained by the majority of other participants. Especially the stated fact is shown in the 2006 expert search task where our system concedes on MAP to only one automatic run.

TABLE I. *OPTIMAL WEIGHTING COEFFICIENT VALUES AND PRECISION FACTORS FOR RUNS*

Run	C_t	C_e	C_b	l	MAP	P@5	P@20
2005q	0.4	0.001	0.57	5	0.1597	0.296	0.203
2005qn	0.4	0.001	0.56	5	0.1464	0.268	0.181
2006q	0.4	0.17	0.51	0.5	0.5929	0.616	0.510
2006qn	0.4	0.1	0.46	5	0.4755	0.522	0.429
2007q	5	0.1	10	5	0.3655	0.192	0.079
2007qn	0.0001	1	0.5	100	0.3622	0.188	0.078
2007ind	ind.	ind.	ind.	ind.	0.5249	0.260	0.105

Optimization of settings showed some reserves for accuracy increase. However, as was discovered subsequently, certain *external* factors relative to our system, such as the possibility to choose a short query, a narrative to it or both fragments together (query+narrative) for expert search, can influence overall system performance capability much more than internal settings. The MAP value for our manual TREC 2007 run where setting parameters and query type were fitted individually for each query (see the last line in Table I) exceeds much the MAP values reached in automatic TREC 2007 runs.

From the examination of our system response on queries we found out that system reaction to different queries differs strongly. Consider TREC 2007 queries with narratives. On Fig. 1 for each query the histogram bar on the left represents the relative (to its average over queries) value of a parameter of system response to terms[1] decreased by 1, and the bar on the right represents the relative number of terms in a query, also decreased by 1. It is quite natural that the system response to terms correlates with query length. From the viewpoint of system response, the TREC 2007 topics turn out to be distinctly divided into two halves. The first 25 queries with narratives are rather short, as a rule. Among them there are 7 topics on which our system is not able to yield the correct answer at any settings, moreover, both on short queries and on queries with narratives. The narratives to these queries are generally characterized by a common-used lexicon abundance or existence of several significant words which are not related to a query subject. The second 25 topics are more “clear” and simple for our system. For the majority of them there is no need to use narratives, and at the same settings (see Table I) the system gives rather precise answers to short queries. But in 5 cases the answer precision increases sharply just at the expense of high-quality narratives. It is clear, that our ranking algorithms will work more precisely with those questions on which the system shows strong response to terms, as in such questions there is more chance to meet the terms characterizing relevant experts.



[1] As such response parameter, we used the combination of query terms significance and their connection cardinality with candidates.

Figure 1. Reaction of the enterprise expert search system to the queries of the CSIRO corpus: left bars are relative values of a response to terms decreased by unity, right bars are relative query-plus-narrative lengths decreased by unity.

It is known from TREC materials [2] that the topics from 26th to 50th are developed by the same CSIRO science communicator. The interesting fact is that exactly these topics proved to be more saturated by terms among all TREC 2007 topics.

As far as the question about the necessity of using narrative is concerned, it should be said that we began a special research of it. We understand that the short query is the theme required experts to be found by user, and therefore, this is the most preferred expert search way. But we also see that to *predict* the necessity for using narrative we can be guided only by some internal calculated system parameters. Terms and bigrams system responses for short queries appear to provide no information about the necessity for using narrative: in some cases the use of narratives leads to improvement of answer accuracy, in other cases – to deterioration of accuracy, regardless of the values of response parameters. There are queries that we suppose to be "difficult" for our system. There are six examples of such queries in Table II. Interesting that even for some of them we can receive quite precise results, but the feature is that the optimal value of average precision (AP_{opt}) is kept in a too narrow range of the setting coefficients, with the coefficients being noticeably different from the shown coefficients in Table I. In other words, the system can provide the right answer to such queries only with nonstandard settings.

TABLE II. EXAMPLES OF "DIFFICULT" QUERIES AND THE CORRESPONDING OPTIMAL VALUES OF SETTING COEFFICIENTS

Query	Type	C_t	C_e	C_b	l	AP_{opt}
13) human clinical trials	qn	10	1	1	1	1
22) airborne hyperspectral	qn	1	0.1	100	10	0.83
40) Southern Surveyor	qn	3	0.1	10	1	0.53
46) recycled water	qn	1	0.1	100	1	0.64
48) polymer bank notes	qn	1	0.1	1	100	1
49) atmosphere	qn	1	1	1	0.5	0.68

It should be mentioned that the nonstandard set of weight coefficients in fact means the domination of one or two expert ranking algorithms. For example, for question № 13 (Table II compared with Table I) the right answer is got by identification of query-candidate connections by terms, for question № 22 it is better to evaluate them with bigrams, and in the question № 49 the great number of the expanding terms should be kept in mind.

So the experimental results give us a weighty reason to suppose that the optimal (in sense of expert search precision) choice of query type for asking the system should be connected to a large extent with its

length, formulation, significance of contained words. We can even talk about some characteristic – the quality of the query, using which 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 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 "narrated" query assessments, on the base of which the most proper ranking variant could be provided to maximize the answer precision. Such two-stage algorithm can highly increase the expert finding precision in thematically wide and general topics or with large number of relevant experts in a particular sphere.

It should be said that we have revealed some inaccuracies in the relevance judgments file mapping. Some experts who are mapped in the TREC 2007 relevance judgments file as relevant were found to be non-existent in the CSIRO collection. Really we have registered several cases of email misprinting, usage of different name forms, as well as one email address belonging to several people. In addition, some addresses from the relevance judgments file do not exist on the web-pages of the CSIRO corpus.

As for implementing the TREC 2007 expert search task in our system, the following fact should be mentioned. The model realized in the system implies a different set of initial data for expert search than in the TREC 2007 collection: our model is adapted to the expert search with email collection, whereas the TREC 2007 documents have only authors and no addressees. Thus we simplified our model for completing the official TREC 2007 expert search task. We can conclude that we applied our model to the simpler task.

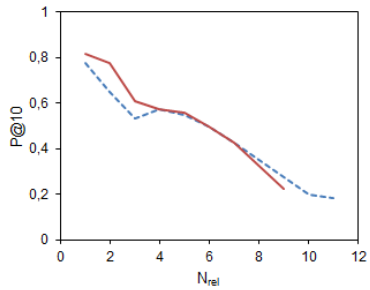


Figure 2. The dependence of the precision at 10th rank on the full amount of query-relevant experts: solid curve – for relevance judgments file mapping without consideration of non-existing experts; dashed curve – for initial relevance judgments file mapping.

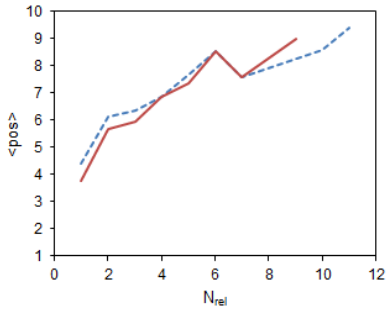


Figure 3. The dependence of the average position of found relevant experts on the full amount of query-relevant experts: solid curve – for relevance judgments file mapping without consideration of non-existing experts; dashed curve – for initial relevance judgments file mapping.

Finally note that in contrast to search precision, the recall has not been quantitatively estimated, but it can be shown indirectly that the large groups of experts in the case of TREC 2007 collection require special treatment for their full detection. We have given the graphics of the relevant experts number among top-10 candidates (Fig. 2) and their average position in the list (Fig. 3) depending on the total number of query-relevant experts. Here, the solid curves have been obtained using the modified relevance judgments file, where we have excluded the non-existing experts from consideration and, consequently, properly reduced the total number of query-relevant experts. And the dashed curves have been obtained using the initial relevance judgments file mapping. As a result, we can assert that the system always finds one or two experts at the top of the ranked list, and other query-relevant experts are usually somewhere deep in the list.

We suppose that this fact is not the system shortcoming for solving the task of finding someone, who knows the topic, but it can become an obstacle when there is the need to find everyone who, for example, knows the details of concrete technology or the production process. In the latter case the great role belongs to the part of our expert search system which gathers the communicative information and which has been successfully implemented to the expert search task in the email corpus W3C.

CONCLUSIONS AND FUTURE EXPLORATIONS

We applied our enterprise expert search system to the official expert search tasks of TREC 2005 – 2007. The model handled these tasks successfully. The model is flexible enough to enable heterogeneous and multilingual collection handling. The search efficiency demonstrated on the English-language W3C and CSIRO corpora is appropriate for practical use of the system and exceeds the efficiency shown by most of TREC participants.

From the viewpoint of search efficiency, we established the optimal weights for the three explored expert ranking algorithms which associate candidates with terms, expanding terms and bigrams. Our algorithms enabled us to detect groups of queries answers to which are high-relevant and stable to the system setting parameters. User's freedom affects reply efficiency in these query groups – as a query the user may choose a short phrase, a text narrating it or both these sources. On the other hand, there also exist queries that cannot be adequately answered in our system, using any available query form. The system shows low relevance on such queries, we say that they are difficult for our system. It is probable enough that other expert search models (such as, e.g., a document-based two-stage model) can reveal more precise results on such queries.

The settings of the described enterprise expert search system could essentially improve search indicators if they are used individually for every query. About a half of all CSIRO queries can be implemented at constant settings, and for reaching high response relevance difficult queries should be handled with deviation from weight coefficient balance, i.e., relying only on one or two from the three described expert ranking algorithms. Here the role of cutting expanding terms by a significance level may be essential for detecting high-significant terms associated with the query.

During the research we attempted to reveal primary signs indicating the measure of query "understandability" in the system. This quality of a query is prescribed by an indexed document collection rather than by internal search engine properties. We suppose query quality being appreciably influenced by such its features as the

number of containing words and especially their significance. If we had some mechanism for estimating query quality or forecasting necessity to specify the query, we could significantly improve the search efficiency. The question about query quality requires further exploration. What is the criterion of a "good" query formulation for the system, how complete must user's information be for querying, how an effective query modification suggestion can 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.

ACKNOWLEDGMENT

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. N. Craswell, A.P. de Vries, I. Soboroff, "Overview of the TREC-2005 enterprise track", in *Proceedings of Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, 2005, pp. 16–22.
2. P. Bailey, N. Craswell, I. Soboroff, A.P. de Vries, "The CSIRO enterprise search test collection", SIGIR Forum, vol. 41, 2007, pp. 42–45.
3. Y. Cao, J. Liu, S. Bao, H. Li, N. Craswell, "A two-stage model for expert search", Technical Report MSR-TR-2008-143, Microsoft Research, 2008.
4. P. Bailey, N. Craswell, A.P. de Vries, I. Soboroff, "Overview of the TREC 2007 enterprise track", in *Proceedings of the 2007 Text REtrieval Conference (TREC 2007)*, Gaithersburg, MD, 2007, pp. 30–36.
5. H. Duan, Q. Zhou, Z. Lu, O. Jin, S. Bao, Y. Cao, Y. Yu, "Research on enterprise track of TREC 2007 at SJTU APEX lab", in *Proceedings of the 2007 Text REtrieval Conference (TREC 2007)*, Gaithersburg, MD, 2007, pp. 489–498.
6. B. He, C. Macdonald, I. Ounis, J. Peng, R.L.T. Santos, "University of Glasgow at TREC 2008: experiments in blog, enterprise, and relevance feedback tracks with Terrier", in *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008, pp. 368–380.
7. J. Zhu, D. Song, S. Rüger, "The Open University at TREC 2007 enterprise track", in *Proceedings of the 2007 Text REtrieval Conference (TREC 2007)*, Gaithersburg, MD, 2007, pp. 431–434.
8. Z. Ru, Q. Li, W. Xu, J. Guo, "BUPT at TREC 2006: enterprise track", in *Proceedings of Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, MD, 2006, pp. 151–156.
9. W. Lu, S. Robertson, A. Macfarlane, H. Zhao, "Window-based enterprise expert search", in *Proceedings of Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, MD, 2006, pp. 186–193.
10. Y. Fu, W. Yu, Y. Li, Y. Liu, M. Zhang, and S. Ma, "THUIR at TREC 2005: enterprise track", in *Proceedings of Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, 2005, pp. 772–779.

11. G. You, Y. Lu, G. Li, Y. Yin, “Ricoh research at TREC 2006 enterprise track”, in *Proceedings of Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, MD, 2006, pp. 570–582.
12. K. Balog, I. Soboroff, P. Thomas, P. Bailey, N. Craswell, A.P. de Vries, “Overview of the TREC 2008 enterprise track”, in *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, Gaithersburg, MD, 2008, pp. 14–25.
13. H. Raghavan, J. Allan, A. McCallum, “An exploration of entity models, collective classification and relation description”, in *Proceedings of the ACM SIGKDD Workshop on Link Analysis and Group Detection*, Seattle, 2004, pp. 1–10.
14. M. Sanderson, *Performance measures used in image information retrieval*, chapter in H. Müller, P. Clough, T. Deselaers, B. Caputo (eds), ImageCLEF: The Information Retrieval Series, vol. 32, 2010, pp. 81 – 94.
15. I. Soboroff, A.P. de Vries, N. Craswell, “Overview of the TREC 2006 enterprise track”, in *Proceedings of Fifteenth Text REtrieval Conference (TREC 2006)*, Gaithersburg, MD, 2006, pp. 32–51.