A New Model for Enterprise Expert Retrieval

Valentin O. Molokanov, Dmitry A. Romanov, and Valentin V. Tsibulsky

***Abstract*—We present a description of an enterprise expert search system which is based on the analysis of content and communications topology in an enterprise information space. As data sources we use the collections introduced at the Text Retrieval Conference (TREC) in 2006 and 2007. 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.**

***Index Terms*—Expert search, TREC, 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.

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 Organisation (CSIRO) enterprise collection [2] which is the crawl of the open-access information from the CSIRO official site.

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.

Our model is essentially candidate-based, that is, we associate candidates with query without primary search of documents on query. As a result, there is no need for us to save whole document texts in the index. But the method of establishing a query-candidate association consists in using some specific techniques, among them are term weighing, building associative connections of a candidate with terms and bigrams, building associative connections between terms, combining several expert ranking ways. In this paper we give the detailed description of our proposed model which we apply to the TREC Enterprise track expert search tasks 2006 and 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.

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. It should be said that there is no conventional expert retrieval approaches for enterprise systems, and expert search models shown at TREC are rather different. Each of them has its own merits, but here we will make a mention of those models which we consider to have definite lacks: they gather information that can be superfluous for expert retrieval. So let us present their basic shortcomings in comparison with our model.

One of widespread expert search models which were carried to an acceptable completion level in 2005 is a document-based two-stage model [3]. Later many TREC participants used some variant of this model [4]. The two-stage expert search model implies two stages in obtaining the resulting expert list: these are document search and people search in relevant documents. Generally speaking, document-based expert search engines require too large index spaces since they must save source documents. Next, notice that the TREC collections contain only open-access information. In our opinion, saving texts may become harmful if some part of information is confidential; here it already depends on the possible extent of information leakage, in particular, disclosure of document texts and personal employee data. Any leakage does not guard the rights of employees and can be inconsistent with law. Such socio-ethical issues were already raised [5]. Thus, in creating

Manuscript received August 10, 2012; revised September 27, 2012. 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".

The authors are with the Science and Education Center of Information Management Technologies, Moscow, Russia (e-mail: vomolokanov@it.ru, dromanov@hse.ru, vtsibulsky@hse.ru).

our model we preferred to calculate terms' statistical properties in the collection, rather than to save analyzed documents fully.

Another group of expert search techniques represents document mapping oriented methods. It can be HTML mapping, email messages mapping by fields, header-based mapping, some bold-facing text strings, etc. ([6], [7], [8]). Indeed, result accuracy is achieved by connecting a candidate with a term from the same structural unit of a document. However, the possibility of work with any documents is essential for us, so a term-candidate association is built due to realistically chosen measures in our model.

Some TREC participants [6] also extracted candidates' expertise information from beyond the collection to improve search efficiency. Again, it does work in the case of open-access information collections, but if the collection were corporate and included commercial secret, finding information on external sources would apparently lose the effectiveness. Our expert search system is developed for corporate collections and always operates in the scope of the collection.

So we propose an alternative approach for expert search as compared to the models listed above. By applying our approach to the TREC 2006 – 2007 expert search task we obtain the result which does not concede to the most of participants. Our model's description is 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.

A. Terms and Bigrams Weighing

For expert search on basis of a set of terms used by employees we introduce a metric which enables to distinguish certain people from other. In our model such a metric is associated with lexicon and a corresponding weight is attributed to each term in the collection. As a natural weight measure we use a significance which we calculate as follows.

Let $n(t, p)$ is the usage amount of a term t by a person p, and $N(p)$ is the total amount of terms used by a person p in the collection. We can find a usage frequency of a term t by a person p:

$$f(t, p) = \frac{n(t, p)}{N(p)} \tag{1}$$

Denote this frequency average logarithm value as $e(t)$, i.e.,

$$e(t) = \frac{1}{P(t)} \sum_{p} \ln f(t, p) \tag{2}$$

where $P(t)$ is the number of employees who used a term t. By marking also the variance of usage t by p frequency logarithm as $D(t)$, namely

$$D(t) = \left[\sum_{p} (\ln f(t, p) - e(t))^2 \right]^{1/2} \tag{3}$$

we define the term significance:

$$S(t) = -\frac{D(t) e(t)}{P(t)} \tag{4}$$

In an analogous way, we also apply the presented approach to bigrams, in this case a bigram should be implied as t in Eqs. (1) — (4).

It is important to emphasize that Eq. (4) is obtained from the most general considerations and it has an easy explanation. A common-used term is characterized by an equal probability of its usage by every employee in a company. The frequency (1) of this term usage by an employee has a log-normal distribution. But if we speak about a professional vocabulary then a term usage probability begins to differ among employees. So the variance (3) is a significance criterion for a term. The term usage frequency average logarithm (2) in the numerator of Eq. (4) as well as the number of employees in the denominator indicate that priority is given to more often used words among lesser number of employees.

Thus, our approach to a term significance metric has quite a feature consisting in detaching common-used rarely occurring words from professional ones, often used in small employee groups. One of significance metrics which do not enable this is, for example, a TFIDF metric. We think that a TFIDF significance definition is appropriate only provided there is no additional information apart from document collection. Undoubtedly, this condition holds for a certain number of tasks. For example, when finding documents in the Internet. However in the expert retrieval case a collection always contains extra information which we use in our significance determination.

The term weighing techniques is applied for some other purposes in [9]. In that work term weight was estimated based on its occurring frequency in high-relevant documents — actually after performing search with an initial query. This weighing was necessary for expanded query generation for each TREC 2007 topic; such a query included a fixed number of the most weighing terms. However before document retrieval there was only an original query with all terms weighing equally. So in our expert search model and [9] the difference between the used term weight estimations proves to be very serious. As opposed to [9], our model assigns to each term its own significance *originally* (at the moment the collection is being indexed). It is also interesting that the term "csiro" is contained in the expanded query in [9] (see Fig. 1 there), but in our model it turns out to be one of the least significant terms in the collection.

B. Term-Candidate and Bigram-Candidate Connections

We estimate employee's topic knowledge level based on corresponding terms usage statistics. One of the key contributions to term-based or bigram-based expert ranking belongs to the term-candidate (bigram-candidate) connection

cardinality. For brevity we will speak about terms below, with meaning that one could imply also a bigram instead of a term.

Now, denote: $n_s(t, p)$ — the total usage number of a term t by a person p; $n_r(t, p)$ — the total usage number of a term t in the messages received by a person p; $c_s(t, p)$ — the number of people who sent a term t to a person p; $c_r(t, p)$ — the number of people who received a term t from a person p. In the case of $n_s, n_r, c_s, c_r \neq 0$ we define the cardinality of a person p's connection with a term t by an expression

$$L(t, p) = \ln n_s(t, p) + \ln n_r(t, p) + \ln c_s(t, p) + \ln c_r(t, p) \quad (5)$$

If any of the parameters n_s, n_r, c_s, c_r is equal to zero then we modify Eq. (5) in such a way that a corresponding logarithm is excluded from there (i.e. so as if the logarithms of zero values were equal to zero).

Thus, the term-candidate association is built not only based on the term usage frequency, but also depending on candidate's topologic features in the subnetwork of the term. The term subnetwork is the graph whose vertices are representing people that have sent or received the term, and the edges are modeling email messages containing this term. Each edge has its weight affected by the number of such messages between two corresponding employees. Depending on the amount of incoming and outcoming edges, each person possesses specific properties in the term subnetwork and these properties add a contribution to the term-candidate connection cardinality.

C. Connections between Terms

The aim of term-to-term connecting in our expert search model is to discover the most significant terms thematically associated to the query terms. It is clear that the more significant is a term, the more a topic knowledge difference reveals between people, so to increase people ranking precision it would be preferable to use significant query terms in the system. It is the significant terms that can "pull" relevant experts to the first positions in the list. The applied terms associating method enables the user to expand a query: typically when writing a query the user hardly ever resorts to using high-significant words, so due to term-to-term connections our model implements the mechanism of detecting the significant lexicon relevant to the query.

We calculate the connection coefficient between two terms t_1 and t_2 as follows. We treat the document text as a sequence of terms. If in any fragment of this sequence the terms t_1 и t_2 are no more than 15 terms distant from each other, then an increment equal to $\frac{1}{\log_2(2+d)}$ is assigned to the connection coefficient between t_1 and t_2, where d is the number of terms situated between t_1 and t_2. The total connection coefficient between t_1 and t_2 is obtained by summing mentioned increments from all text fragments in which there are no more than 15 terms between t_1 and t_2.

So 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 [10].

D. Expert Ranking Algorithms

Our expert search model uses three expert ranking algorithms. Each of them yields candidate's connection coefficient with query terms, expanding terms and bigrams respectively. The final rank of a person p relative to a query is calculated as

$$W(p) = C_t W_t(p) + C_e W_e(p) + C_b W_b(p) \quad (6)$$

where $W_t(p)$ is candidate p's direct connection coefficient with query terms, $W_e(p)$ is candidate p's indirect connection coefficient with expanding terms (i.e. terms associated with query terms), $W_b(p)$ is candidate p's direct connection coefficient with bigrams contained in the query, and C_t, C_e, C_b are corresponding setting parameters that could be specified by a user. As can be seen from (6), the system setting parameters are actually those weighting coefficients that combine the expert search algorithms in consideration. Below we present calculation algorithms for each of the three connection coefficients appearing in Eq. (6).

Let T is a set of unique terms in a query. Candidate's direct connection coefficient with query terms is defined by an expression

$$W_t(p) = \sum_{t \in T} S(t) L(t, p) \quad (7)$$

where $S(t)$ is the term t's significance calculated in Eq. (4), and $L(t, p)$ is candidate p's connection cardinality with the term t defined by Eq. (5).

Candidate's direct connection coefficient with bigrams is defined similarly as in Eq. (7). Let B is a set of unique bigrams in a query. If a bigram b is implied in Eqs. (4) and (5) instead of t then candidate's direct connection coefficient with bigrams is determined as

$$W_b(p) = \sum_{b \in B} S(b) L(b, p) \quad (8)$$

Candidate's indirect connection coefficient with expanding terms is calculated in some different way. Namely, consider a set of expanding terms T' containing all collection terms which are not present in a query but are associated with any query term. For each expanding term $t' \in T'$ we calculate a coefficient

$$X(t') = \sum_{t \in T} S(t) S(t') \ln C(t, t') \quad (9)$$

which could be described as a connection weight of this term with a query; here $C(t, t')$ is the connection coefficient

between the terms t and t' computed according to the procedure described in Section 3.3. For further consideration from all expanding terms we select the terms that are *strongly* associated with the query by specifying one extra setting parameter. For such terms, the coefficient (9) exceeds its average value by a given cutting level. Concretely, if we label the value (9) averaged over all expanding terms as

$$\overline{X} = \frac{1}{N}\sum_{t' \in T'} X(t') \tag{10}$$

(here N is the total expanding terms amount) and the value (9) variance as

$$\sigma = \left[\frac{1}{N}\sum_{t' \in T'} (X(t') - \overline{X})^2\right]^{1/2} \tag{11}$$

then an expanding term t' is thought to be strongly associated with the query if the condition

$$\frac{X(t') - \overline{X}}{\sigma} > l \tag{12}$$

holds for this term; here l is a specified cutting level being set by a user.

Thus, all expanding terms providing the condition (12) implementation form a set E of terms strongly associated with the query, and only such terms $e \in E$ take part in calculating candidate's indirect connection coefficient with expanding terms:

$$W_e(p) = \sum_{e \in E} S(e)L(e, p) \tag{13}$$

So in the description given above we have presented expert ranking algorithms used in our enterprise expert search system. People connection coefficients with query terms, expanding terms and bigrams are obtained with those algorithms and represent three distinct people ranking parameters. By combining these ranking parameters taken with corresponding weights set by a user, our model calculates the final expert's rank which affects expert's position in the list.

Note that we realized the described model in two forms: base and modified ones. They differ by the total amount of terms stored in the system and the modification consists in expanding the set of terms. Apart from the terms contained in the base model, our modified model also finds emails and phone numbers and saves them as separate terms. As a result, if an email or a phone number is connected with a person, then this person gains additional associations with the terms mentioned near his contact information, according to our algorithms. The results show that this modification increases the overall efficiency by optimizing expertise fields retrieval from people's contact details.

IV. Results

To compare and optimize expert search results, we performed a series of system runs with different settings. We changed both the weighting coefficients for the considered lexical types of expert ranking (query terms, expanding terms, bigrams) and the number of expanding terms involved in calculations. For each run we fixed the values of search precision metrics accepted on TREC: these are mean average precision (MAP), precision at 5th (P@5) and 20th (P@20) ranks [11].

We found such sets of settings which are constant for all queries of a collection and give the best MAP value in comparison with other possible settings options (Table I); with these settings, other considered precision factors also appear near an optimum. Comparing the results of our base and modified model runs on the 2006 and 2007 corpora with TREC participants' results (see, respectively, Table IV in [12] and Table IV in [4]), it is possible to conclude that our shown expert search accuracy surpasses the accuracy obtained by the majority of other participants. Interestingly, that in the 2006 expert search task our modified model exceeds all other automatic runs in MAP.

TABLE I: Optimal Weighting Coefficient Values and Corresponding Precision Factors

Run	C_t	C_e	C_b	l	MAP	P@5	P@20
2006q	0.4	0.17	0.51	0.5	0.593	0.616	0.510
2007q	5	0.1	10	5	0.366	0.192	0.079
2006qMod	0.4	0.17	0.51	0.5	0.595	0.620	0.513
2007qMod	1.3	0.01	10	2.8	0.392	0.208	0.081

Although the W3C and CSIRO collections have much in common, let us pay some attention to the differences in their query nature and in the corresponding answer quality. While viewing query texts, we can notice that there is more specialized lexicon in the W3C queries in general, as compared with the CSIRO queries. The same is really reflected in the calculated significance of words: the W3C query terms proved to be more significant. We recognized that the system yields a quite accurate (with average precision not less than 0.3) answer to almost any W3C query, while in the CSIRO corpus such AP values are achieved approximately in a half of queries. On the other hand, we obtained several answers with AP=1 in the CSIRO collection. This is due to a lesser number of mapped experts, since the 2007 task was specially formed towards engines with high early precision rather than recall [13]. We conclude that our model is essentially precision-oriented as well: as a rule, our answers to the 2007 queries contain one or two relevant experts in the top of the list and other relevant experts somewhere in the depth of the list.

To summarize, our proposed model always provides high expert search precision, and in the case of a specialized corpus and significant query terms a nice recall is also provided. We treat our model to be especially convenient for corporate environments.

V. Conclusions and Further Explorations

We applied our proposed model to the official expert search tasks of TREC 2006 – 2007. The model handled these tasks successfully and outperformed most TREC participants' models. The model is flexible enough to enable heterogeneous and multilingual collection handling.

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. Emphasize that using the same index our model allows to employ several expert rating factors in the system as well as vary their combinations, i.e., create a final metric as a derivative from them. This is our model's development specificity.

Notice that we successfully chose those heuristic indicators which model a query-candidate association. Our methods give us a quite high early precision without requiring to save source documents in the expert search system resources or to attract any data from beyond the collection. In addition, if the lexicon of a collection is specialized and the query terms are significant enough, our model also provides an excellent recall. So for corporate media our expert search system is particularly suitable.

Finally, our model reveals additional improvement possibilities. First, the model has reserves in algorithms, as any new introduced expert ranking algorithm can be combined with the considered ones. Second, there exist other noticeable performance reserves in the model. If we construct a more or less universal metric which could evaluate answer precision based on some parameters calculated in the system without using the relevance judgments file, we will be able to realize some mechanism for estimating query "understandability" in the system. For example, if a query turns out to be "difficult" for the system, then the system could offer a user to specify this query and perform expert search on the specified query, giving a more precise answer. The issue about query understandability in the system 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 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.

REFERENCES

[1] N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the TREC-2005 enterprise track," *TREC 2005*, Gaithersburg, 2005, pp. 16-22.

[2] P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries, "The CSIRO enterprise search test collection," *SIGIR Forum*, vol. 41, no. 2, pp. 42-45, Dec. 2007.

[3] Y. Cao, J. Liu, S. Bao, H. Li, and N. Craswell, "A two-stage model for expert search," *Technical Report MSR-TR-2008-143*, Microsoft Research, Oct. 2008.

[4] P. Bailey, N. Craswell, A. P. de Vries, and I. Soboroff, "Overview of the TREC 2007 enterprise track," *TREC 2007*, Gaithersburg, 2007, pp. 30-36.

[5] S. Lichtenstein, S. Tedmori, and T. Jackson, "Socio-ethical issues for expertise location from electronic mail," *Int. J. Knowledge and Learning*, vol. 4, no. 1, pp. 58-74, Mar. 2008.

[6] Y. Fu, W. Yu, Y. Li, Y. Liu, M. Zhang, and S. Ma, "THUIR at TREC 2005: enterprise track," *TREC 2005*, Gaithersburg, 2005, pp. 772-779.

[7] Z. Ru, Q. Li, W. Xu, and J. Guo, "BUPT at TREC 2006: Enterprise track," *TREC 2006*, Gaithersburg, 2006, pp. 151-156.

[8] G. You, Y. Lu, G. Li, and Y. Yin, "Ricoh research at TREC 2006 enterprise track," *TREC 2006*, Gaithersburg, 2006, pp. 570-582.

[9] K. Balog and M. de Rijke, "Non-local evidence for expert finding," in *Proc. of 17th ACM Conf. Information and Knowledge Management*, New York, 2008, pp. 489-498.

[10] H. Raghavan, J. Allan, and A. McCallum, "An exploration of entity models, collective classification and relation description," *ACM SIGKDD Workshop on Link Analysis and Group Detection*, Seattle, 2004, pp. 1-10.

[11] M. Sanderson, "Performance measures used in image information retrieval," *Experimental Evaluation in Visual Information Retrieval*, H. Müller, P. Clough, T. Deselaers, and B. Caputo, Ed. New York: Springer, 2010, ch. 5, pp. 81-94.

[12] I. Soboroff, A. P. de Vries, and N. Craswell, "Overview of the TREC 2006 enterprise track," *TREC 2006*, Gaithersburg, 2006, pp. 32-51.

[13] P. Bailey, D. Agrawal, and A. Kumar, "TREC 2007 enterprise track at CSIRO," *TREC 2007*, Gaithersburg, 2007, pp. 205-210.