Expertise Retrieval Using Search Engine Results

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

Shuguang Han
Center for Studies of Information Resources, Wuhan University
Wuhan, P. R. China
hanshuguang@gmail.com

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

ABSTRACT

Expertise retrieval has been largely explored on a few collections crawled from the intranets of organizations. In contrast, only limited external information has been used and studied. In this paper, we have a research on the approaches and effectiveness of expertise retrieval using search engine results. Using appropriate queries, we search for each expert his or her relevant information from the internet and create collections that are quite different from the intranet ones. On such basis, different search queries are compared for the effectiveness of their results. Further, we try on different fields of the results and make a comparison between their effects. Besides, results inside and outside the organization are experimented separately to make clear their different effects. In our experiments, the language modeling approach of expertise retrieval still works well with search engine results. To conclude, we suggest that search engine is an effective source of expertise information and can produce considerable performance in expertise retrieval.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [**Information Systems Applications**]: H.4.2 Types of Systems; H.4.m Miscellaneous

General Terms

Measurement, Performance, Experimentation

Keywords

Expertise retrieval, expert finding, search engine results

1. INTRODUCTION

Expert finding has been studied sporadically since the late 20th century, but not been highly focused on before the expert search task appeared in TREC. Yimam-Seid identified two main motives in expert finding [1], i.e. the information need and the expertise need. The latter motive refers to the object of finding experts with given expertise, which has become the main foci of the TREC expert search task in the past three years. In this paper, we also concentrate on the latter motive. For convenience, all the occurrences of the notion *expert finding* and *expertise retrieval* in this paper refer to the latter motive specifically.

In the TREC expert search task, expertise retrieval is explored mainly on the basis of a few intranet collections. These collections consist of heterogeneous information inside the organizations, aiming at simulating real information needs. Though disputed on a few problems, the TREC collections have largely facilitated researches on expertise retrieval and brought some effective and robust formal models. In contrast, only limited external information has been used and studied. Though it is natural for us to consider that the organization itself should hold more information than anyone else, without any substantial evidence, the external information cannot be overlooked.

As a result, we have a research on the approaches and effectiveness of expertise retrieval using search engine results. In our experiments, the intranet collection is only used for extracting a candidate list of the organization.

Our foci in this paper involve the following problems: first, what is the difference between the intranet collection and the search engine results; second, how to generate and make use of the search engine results effectively; third, whether the language modeling approach of expertise retrieval still works well in such circumstance; fourth, in the search engine results, whether the results inside the organization are more effective than those outside; fifth, whether the intranet collection can overwhelm the search engine results in effectiveness.

The remainder of this paper is organized as follows. In section 2, we have a discussion on the use of external information in expertise retrieval. Section 3 explains the approaches and models of expertise retrieval using search engine results. In section 4, some details of our experiments are introduced. Section 5 evaluates the results of experiments and answers the problems we proposed. In section 6, we draw a conclusion from our research and propose some future challenges.

2. USING EXTERNL INFORMATION

Existing collections for expertise retrieval mainly consist of information from the intranets of organizations. In the past three years, the TREC expert search task has provided two intranet collections, i.e. the W3C collection [2] and the CERC collection [3]. Besides, Balog et al. have created the Uvt collection [4], which is comparatively small collection comprising bilingual (English and Dutch) information crawled from Tilburg University.

In contrast, only limited external information is used and studied. Some general supplementary methods are often used in expertise retrieval, which sometimes involve the use of external information. For example, Troy et al. adopted the WordNet to identify synonyms for query expansion [5]. Though effective, such resources contains hardly any expertise information and thus are not focused on in this paper.

Future Challenges in Expertise Retrieval
SIGIR 2008 workshop, July 24, Singapore

What we are interested in are the resources that can provide much expertise information. Generally, there are two kinds of such resources, i.e. the general search engine and the specialized database.

The general search engine crawls for information from the internet, which involves data both inside and outside the organizations and covers various formats of resources. Given appropriate queries, the search engine can return ranked results relevant to the experts, which can be extracted and gathered automatically to aid the expertise retrieval.

Some specialized database can also provide important expertise information. For example, the academic database provides information about the literature of experts, which can be used to evaluate their expertise. Besides, the patents invented by experts can also be used for judging expertise, thus the patent database will also facilitate the expertise retrieval. But most of the specialized databases charge for service, which makes relatively higher costs to access information.

In recent years, vertical search [6] has been largely advanced and the vertical search engine may act as an effective substitute of the specialized database. Compared with the specialized database, the vertical search engine, which, in contrast, is almost free of charge, can provide integrated results covering a large amount of the specialized databases and web pages.

Among the TREC participants, Chu-Carroll et al. have ever used information from Google Scholar to assist expertise retrieval [7]. In their work, an author list is generated and extracted by searching the query in Google Scholar. Besides, the publications and the citations are also extracted and recorded. In the end, the author's expertise is computed considering the quantity of publications, citations of each publication and the author's position in the author list. However, in their work, only results of the method that integrated Google Scholar and the intranet collection are provided. As a result, no comparison is available for these two kinds of information.

In this paper, we mainly focus on the issue that whether it is feasible to retrieve experts on the basis of the external resources rather than the intranet collections. Search engine is used to explore this problem: firstly, it provides integrated information both inside and outside the organization; secondly, it is publicly available and free of charge; besides, the various formats of queries supported by the search engine can help us investigate on some problems further. In the next section, we illustrate the approaches and models of expertise retrieval using search engine results.

3. APPROACHES AND MODELS

In this section, we will illustrate the approaches and models of expertise retrieval using search engine results. The whole process mainly involves the following steps: firstly, a candidate list of the organization should be extracted; on such basis, appropriate queries can be built for each expert to generate relevant results from the search engine; then, useful contents of the results should be extracted and indexed; in the end, experts will be ranked on the basis of the scoring models. The rest of this section will explain these steps in turn.

3.1 Extraction of Candidate List

The candidate list is a listing of experts and their evidence. It provides information to recognize experts in the collection. The expert evidence often involves different variations of person names and email addresses. Considering full name can be used to generate other variations of person names, the candidate list should at least provide the full name and email address for each expert.

The extraction of candidate list can be implemented as a part of a named entity recognition process. Some useful information often helps the recognition, e.g. the email address in the organization often conforms to *firstname.lastname@domain*, which can largely facilitate the recognition process [8]. Besides, the organizations usually provide introductory pages that list its employees. The extracting of candidate list will be largely shortened if these pages can be recognized and analyzed specifically.

In our approaches, the extraction of candidate list is implemented using a rule-based named entity recognition method, which is similar to Mikheev et al. [9]. If the intranet collection is used as the main source of expertise information, the experts should also be located for their occurrences in each document. Since the recognition process is not the focus in this paper, we do not go further here. Some simple evaluation of the recognition is given in section 4.

3.2 Building Search Queries

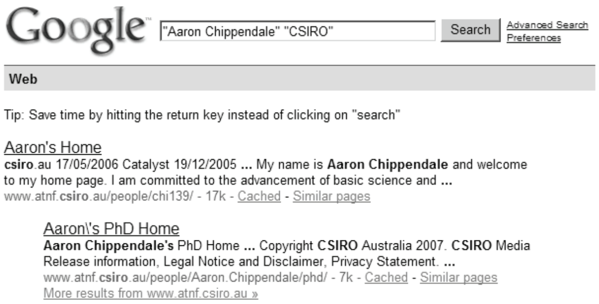
When a candidate list is extracted, we can use the listed evidence of experts to build appropriate queries in order to search for information relevant to the experts from the search engine. But the search query should be delicately designed to generate as many relevant results as possible and avoid non-relevant ones.

For most of the time, searching with the full name can successfully match the expert in relatively small collection, e.g. the internal collection. But for the search engine, which involves huge information all over the internet, only using the full name as query may produce too much noise. As a result, we adopt the combination of full name and the organization name using relation **AND**, namely Q_1, to reduce the effects of name ambiguity. Further, the email address, namely Q_2, can correctly match experts at nearly all occasions, but may be deficient in recall. Besides, Q_1 and Q_2 can be combined using relation **OR**, namely Q_3. Table 1 gives a glance at the basic queries adopted in our approaches.

Table 1. Basic Queries and the corresponding formulas

Query	Formula
Q_1	"Full Name" **AND** "Organization Name"
Q_2	"Email Address"
Q_3	Q_1 **OR** Q_2

Besides, almost all of the general search engines support some extensive functions. Firstly, returned results of search engine are usually filtered as default, which clusters results from the same domain and returns only the most relevant pages rather than all the pages in the domain. Figure 1 gives an example of the filtered results returned by Google. It is showed that only two of the results from the domain atnf.csiro.au are listed, with an access to show more results from this domain. Though possibly improving the searching experience of users, the filter function may result in negative effects for the expertise retrieval. In section 5, we discuss this problem by comparing effectiveness of the three queries with those of queries using the filter function, namely *Q1F*, *Q2F* and *Q3F*. Secondly, it can be restricted to only return results inside or outside the specific domain, which can be used to investigate on the fourth problem proposed in section 1.



"Aaron Chippendale" "CSIRO"

Tip: Save time by hitting the return key instead of clicking on "search"

Aaron's Home
csiro.au 17/05/2006 Catalyst 19/12/2005 **...** My name is **Aaron Chippendale** and welcome to my home page. I am committed to the advancement of basic science and **...**
www.atnf.**csiro**.au/people/chi139/ - 17k - Cached - Similar pages

Aaron\'s PhD Home
Aaron Chippendale's PhD Home **...** Copyright **CSIRO** Australia 2007. **CSIRO** Media Release information, Legal Notice and Disclaimer, Privacy Statement. **...**
www.atnf.**csiro**.au/people/Aaron.Chippendale/phd/ - 7k - Cached - Similar pages
More results from www.atnf.csiro.au »

Figure 1. An example of filtered results returned by Google.

3.3 Gathering Search Engine Results

When results are generated by the search engine, the result pages can be crawled to obtain and store the results. The returned result generally involves the following fields: title, abstract, URL and cache URL. The title field comes from the metadata of the results, e.g. the content within the <title> tags of the web pages. The abstract field is generated automatically to give a glance at the result, which presents the co-occurrences of keywords within an appropriate window size. The URL field gives out the original URL of the result, while the cache URL field links to an archive of the result, which is stored in servers of the search engine.

Different fields of the results may be varied in effectiveness. As a result, we have try on different fields in expertise retrieval. In our approaches, the simple combination of the title and abstract fields, namely *ABS*, are extracted and indexed. Besides, the entire content of the resource can provide complete information, which is also tested, namely *CO*. For some of the results, the entire content may be unable to acquire from the provided URL, e.g. the result is removed from the server. At this time, the content of the cache location will be used instead. Table 2 shows the different contents of the search engine results adopted in our experiments.

During this step, it generates for each expert a list of relevant results, which are stored and indexed as documents and will be used for further scoring step.

Table 2. Different fields of the search engine results

Field	Explanation
ABS	Combination of title field and abstract field
CO	The entire content of resource

3.4 Scoring Models

In the TREC expert search task, some modeling approaches are proposed and proved to be effective and robust. However, these models are testified only within the internal collections. In this paper, one of the main foci is to examine the effectiveness of the expertise retrieval model using search engine results.

The widely adopted model of expertise retrieval follows a language modeling approach, which transforms the problems of assessing relevance between an expert *e* and the query *q* into the estimation of the probability that *e* can be generated by *q*, i.e. $p(e|q)$. According to Bayes Formula, the problem can also be focused on the estimation of $p(q|e)$, if we assume an equal probability for all the experts. Balog et al. discussed two different processes of the language modeling approach in expertise retrieval, i.e. the candidate model and the document model [10]. In the experiments, the latter model can produce better effectiveness than the former one. Assuming conditional independence between the query and the expert, $p(q|e)$ can be estimated as formula 1.

$$p(q \mid e) = \sum_{d \in D_e} p(q \mid d) p(d \mid e) \tag{1}$$

In formula 1, D_e refers to the set of documents containing evidence of *e*; $p(q|d)$ is estimated as a general language model using the Jelinek-Mercer smoothing [11], which is given as the formula 2; $p(d|e)$ can be estimated as the association between the document and the expert, which, for simplification, is set to 1 if *d* is the search engine results return by search *e*. For the intranet collection, $p(d|e)$ is set to 1 if d contains occurrence of full name or email address of e.

$$p(q \mid d) = \prod_{t_i \in q} [(1-\lambda) p_{ml}(t_i \mid d) + \lambda p(t_i \mid C)] \tag{2}$$

In formula 2, t_i refers to each term of the query *q*; $p_{ml}(t_i|d)$ refers to the maximum likelihood estimate of t_i in *d*; *C* is the whole corpus; $p(t_i|C)$ is the probability of t_i in *C*; λ is a smoothing variable which is set to 0.5 in our approaches.

In the scoring process, the content of search engine results will be processed as documents. Then, experts will be ranked according the scoring model. Section 4 explains our experiments in detail and section 5 shows the evaluation of our approaches.

4. EXPERIMENTS

Our main foci in this research involve five problems, which have been presented in section 1. Accordingly, several sets of experiments are set up in order to investigate on these problems.

For the first problem, we collect statistics on the resources of the search engine and compare with the intranet collection. Detail of the statistics is given in section 5. For the second problem, we try on different queries to generate results and make use of different fields of the results to rank experts and compare their effectiveness. For the fourth problem, search engine results will be distinguished by URL to make a comparison between results inside and outside the organization. For the fifth problem, results of previous experiments will be compared with those retrieved from the intranet collection. The answer to the second problem will be shown in the whole process of experiments.

We adopt the CERC collection to conduct our experiments for the following concerns. Firstly, compared with the W3C collection, the candidate list is not officially provided in the CERC collection, which can better simulate the real circumstance. Secondly, the evaluation of the W3C collection is arguable. For TREC 2005, it is not assured whether the experts outside the working groups can be excluded from consideration [12]; for TREC 2006, the evaluation of the pooled results may lack effectiveness to estimate results obtained from the search engine, because the pooled results only include results obtained from the internal corpus [2]. In contrast, the evaluation of the CERC collection is implemented without the pooling method.

Using the recognition approaches described in subsection 3.2, we have extracted a candidate list with 3229 experts from the CERC collection. Though no official result of the candidate list is available, some statistic can reflect the effectiveness: among the 152 experts provided as the relevant experts for 50 topics, 129 experts are recognized and listed in the candidate list.

The CERC collection is processed and retrieved for experts as the baseline experiments and the evaluation is shown in Table 3.

Table 3. Evaluation of baseline results by the CERC collection

run	rel-ret	map	R-prec	P5	P10
baseline	97	0.3823	0.3619	0.216	0.136

As for the search engine, we choose Google as the general search engine in our experiments for its great popularity and quick response. Search engine results are crawled by customized crawler programs. Lucene is used in our experiments to complete most of the indexing of documents.

5. EVALUATION

In this section, a few sets of experiments are evaluated under the CERC collection, aiming at answering the concerned problems proposed in section 1. Firstly, the search engine results are collected and compared with the intranet collection. Secondly, different queries will be experimented on their effectiveness. Then, different fields of the results are also examined. Further, a comparison is made between the results inside and outside the organization. In the end, the comparison is made between the proposed approaches using search engines and other approaches. The answer of these questions will be given through the evaluation of results.

5.1 Search Engine Results

In this subsection, results of each query will be shown and compared with the CERC collection. Among the 370715 documents in the CERC collection, 64416 documents contain expert evidence.

Table 4 gives some statistics on the search engine results for each query. It is revealed that *Q1* can return more results than *Q2*. When *Q1* and *Q2* are combined together, the most results are returned. Besides, the filter function of the search engine can distinctly reduce the returned results for each query. Further, we can conclude that almost for all queries, results outside the csiro.au domain are distinctly more than those inside the domain.

Theoretically, results of Q_1F and Q_2F will be included within the results of Q_1 and Q_2, but practically the statistics does not fully accord with the expectation. Among results of *Q1F*, 100036 results (97.43%) are included in results of *Q1*; as for results of *Q2F*, 49460 results (84.69%) are included in *Q2*. Still, most of the filtered results are included in results returned by normal queries.

For any query, it is clear that most of the results returned are not involved in the CERC collection. To some extent, it can prove that search engine results contain information quite different from that of the intranet collection, which replies to the first question in the introduction section.

Table 4. Evaluation of results using different queries

query	unique results		results included in the CERC collection
	internal	external	
Q_1	75580	184730	19869
Q_1F	14887	93713	4272
Q_2	33862	68817	9506
Q_2F	8228	50171	2836
Q_3	83908	212320	21111
Q_3F	18795	121901	5391

5.2 Effectiveness of Different Queries

In this subsection, we are mainly concerned on the effectiveness of different search queries. Q_1, Q_2 and Q_3 will be compared. Table 5 shows the effectiveness of expertise retrieval using results returned by different queries. The fields used in the results are *CO*. It is revealed that the filter function will impede the effectiveness of expertise, because Q_1F, Q_2F and Q_3F produce relatively lower effectiveness than Q_1, Q_2 and Q_3 do. For the basic queries, Q_1 is much more effective than Q_2. Q_3 is the combination of Q_1 and Q_2 using relation **OR** and performs slightly worse than Q_1. It showed that *Q2* is noisy and somewhat redundant, since it cannot improve *Q1* and also performs the worst itself.

Table 5. Evaluation of results using different queries

run	rel-ret	map	R-prec	P5	P10
baseline	97	0.3823	0.3619	0.216	0.136
Q_1	98	0.3769	0.3199	0.196	0.134
Q_1F	97	0.3615	0.3235	0.196	0.128
Q_2	81	0.1935	0.1600	0.120	0.074
Q_2F	73	0.1592	0.0918	0.090	0.069
Q_3	98	0.3742	0.3165	0.192	0.132
Q_3F	93	0.3559	0.3289	0.204	0.122

5.3 Effectiveness of Different Contents

In subsection 5.2, Q_1 is proved to be much more effective than Q_2 and Q_3 using content *CO*. As a result, we continue using Q_1 and Q_1F to investigate on the effectiveness of different fields. The field *ABS* and CO will be examined under the query Q_1 and Q_1F, i.e. Q_1ABS and Q_1FABS.

Table 6 shows the effectiveness of expertise retrieval using field *CO* and *ABS* separately under the query Q_1 and Q_1F. The negative effects of the filter function can be testified again in the comparison between Q_1ABS and Q_1FABS. *ABS* results in better results under Q_1, but worse results under Q_1F.

Compared with *CO*, *ABS* can enhance the precision of expertise retrieval, in a cost of reducing recall. For both Q_1 and Q_1F, *ABS* returns distinctly less experts, but does not reduce much in map. In fact, the Q_1ABS even receives higher map than Q_1.

Now, the second question proposed in section 1 can be replied in this subsection, that is, using Q_1 without using filter function will produce the most effective queries for expertise retrieval. Comparatively, using *CO* involves more noises but can also enhance recall, while *ABS* involves fewer noises and higher precision but also lower recall.

Table 6. Evaluation of results using different contents

run	rel-ret	map	R-prec	P5	P10
baseline	97	0.3823	0.3619	0.216	0.136
Q_1	98	0.3769	0.3199	0.196	0.134
Q_1ABS	90	0.3953	0.3572	0.225	0.133
Q_1F	97	0.3615	0.3235	0.196	0.128
Q_1FABS	83	0.3533	0.3145	0.196	0.120

5.4 Effectiveness of Different Domains

In this subsection, we focus on the comparison between results inside and outside the organization. The results returned without any domain restrict are integrated information comprising resources both inside and outside the organization. The internal information and the external information can be accessed directly from the search engine using some domain-restrict queries, e.g. "*site: domain*" or "*-site: domain*" in Google, or indirectly by distinguishing from the integrated information according to the URLs. Run Q_1, Q_1F, Q_1ABS and Q_1FABS are all tested for resources inside and outside the organization. Each of them generates two runs, which are distinguished in run names using suffix "*I*" or suffix "*O*" to represent for using results inside and outside the organization. Table 7 shows the evaluation of results from different domains.

Table 7. Evaluation of results from different domains

run	rel-ret	Map	R-prec	P5	P10
baseline	97	0.3823	0.3619	0.216	0.136
Q_1	98	0.3769	0.3199	0.196	0.134
Q_1I	96	0.3985	0.3509	0.212	0.134
Q_1E	90	0.3173	0.2574	0.168	0.114
Q_1F	97	0.3615	0.3235	0.196	0.128
Q_1FI	92	0.3214	0.2689	0.164	0.112
Q_1FE	87	0.3068	0.2650	0.180	0.118
Q_1ABS	90	0.3953	0.3572	0.225	0.133
Q_1ABSI	82	0.3714	0.3002	0.204	0.128
Q_1ABSE	80	0.3360	0.3109	0.192	0.115
Q_1FABS	83	0.3533	0.3145	0.196	0.120
Q_1FABSI	65	0.2667	0.2332	0.156	0.089
Q_1FABSE	78	0.3200	0.2880	0.179	0.109

In table 7, it is revealed that, for most of the occasions, results inside the organization are distinctly more effective than those outside, with the only exception for Q_1FABS. However, the integrated results are much more effective than the separated two results in most of the occasions except Q_1. These conclusion answers the fourth question proposed in the section 1. It should be noticed that for most of the time results inside and outside the organization performs as complements for each other.

5.5 Comparison to Other Approaches

Compared with the baseline run, whose effectiveness is shown in section 4, two runs using search engine results exceed the baseline run, i.e. Q_1I and Q_1ABS. Besides, Q_1 and Q_1ABSI also produce comparable performance. These evaluation results can reply to the fifth question proposed in section 1. Generally, the expertise retrieval approaches using search engine results are examined to be fruitful. It is revealed that the external information, at least search engine, can also contribute effectively to the expertise retrieval. As for the third question, the language modeling approach is testified to be effective with search engine results.

The top ranked results in our experiments are also fruitful when compared with results from other researchers. According to [13], the top four runs in our experiments would be in the top 5 among all the automatic runs in the TREC 2007 expert search task.

6. CONCLUSION

In this paper, we have a research on expertise retrieval using search engine results rather than the intranet collection. In our approach, search engine is used as the main source of expertise information, which is effective and can result in even better results than the intranet collection does in some occasions. Our experiments prove that the external sources of expertise information cannot be excluded from consideration in the expertise retrieval. Different search queries and fields of the results are also examined for their effectiveness. Besides, results inside and outside the organization are experimented separately and compared, which reveals that results inside the organization are generally more effective, but the integrated results can perform the best. A somewhat surprising result is that search engine results are quite different from the intranet collection.

In the future, we suggest that more kinds of the external expertise information should be used and studied in expertise retrieval. What this paper discussed is only one of the external sources of expertise information. Except for the general search engine, some specialized databases and the vertical search engines may also provide important clues for us to improve the expertise retrieval, which should be included in future research.

7. ACKNOWLEDGMENTS

We thank the four anonymous reviewers for helpful comments on this work. This research was supported by National Social Science Foundation of China 06CTQ006.

8. 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] I. Soboroff, A. de Vries and N. Craswell. Overview of the TREC 2006 enterprise track. In Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), 2006.

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

[4] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke and A. van den Bosch. Broad expertise retrieval in sparse data environments. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, Netherlands, 2007, pp. 551-558.

[5] A. Troy and G. Zhang. Case Western Reserve University at the TREC 2006 enterprise track. In Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), 2006.

[6] Wikipedia. Vertical search, 2008. URL: http://en.wikipedia.org/wiki/Vertical_search. Visited at: 20 May 2008.

[7] J. Chu-Carroll, G. Averboch, P. Duboue, D. Gondek, J. Murdock and J. Prager. IBM in TREC 2006 enterprise track. In Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006), 2006.

[8] TREC 2007. TREC-2007 enterprise track guidelines, 2007. URL: http:// www.ins.cwi.nl/projects/trec-ent/wiki/index.php /TREC_2007#Expert_Search_Task. Visited at: 20 May 2008.

[9] 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.

[10] K. Balog, L. Azzopardi and M. de Rijke. Formal models for expert finding in enterprise corpora. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, USA, 2006, pp. 43-50.

[11] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems, 22(2):179-214, 2004.

[12] N. Craswell, A. de Vries and I. Soboroff. Overview of the TREC 2005 enterprise track. In Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005), 2005.

[13] P. Bailey, N. Craswell, A. de Vries and I. Soboroff. Overview of the TREC 2007 enterprise track. In Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007), 2007.