University of Twente at the TREC 2008 Enterprise Track: Using the Global Web as an expertise evidence source

Pavel Serdyukov, Robin Aly, Djoerd Hiemstra
University of Twente
PO Box 217, 7500 AE
Enschede, The Netherlands
{serdyukovpv, alyr, hiemstra}@cs.utwente.nl

ABSTRACT

This paper describes the details of our participation in expert search task of the TREC 2007 Enterprise track.

1. INTRODUCTION

This is the fourth (and the last) year of TREC 2007 Enterprise Track and the second year the University of Twente (Database group) submitted runs for the expert finding task. In the methods that were used to produce these runs, we mostly rely on the predicting potential of those expertise evidence sources that are publicly available on the Global Web, but not hosted at the website of the organization under study (CSIRO). This paper describes the follow-up studies complimentary to our recent research [8] that demonstrated how taking the web factor seriously significantly improves the performance of expert finding in the enterprise.

2. EXPERTISE EVIDENCE ACQUISITION FROM THE GLOBAL WEB

One could imagine an expert finder that is equipped with a web crawler focusing on retrieval of employee-specific information from the Web. Such a spider would provide us with a plenty of information about how the organization is positioned at the world or regional markets, how influential and wide-spread its organizational knowledge. However, in case when an expert finder should be made cheap but good, the enterprise may rely on powerful mediators between people and the Web: leading search engines and their public search APIs.

In our latest studies [8] we found that extracting topic- and person-specific information with Yahoo! and Google Search APIs is a universal way to expand the search scope of expert finders. We used as many expertise evidence sources as possible to finally aggregate ranks from several source-specific rankings per each candidate. We relied on the hypothesis that real experts should be popular not only locally, in the enterprise, but also in the other web spaces available for search: news, blogs, academic libraries etc. We extracted expertise evidence from search engines by issuing queries for each candidate containing:

- the quoted full person name: e.g. *"tj higgins"*,
- the name of the organization: *csiro*,
- query terms without any quotes: e.g. *genetic modification*,
- the directive prohibiting the search at the organizational web site: *-inurl:csiro.au*.

Adding the organization's name was important for the resolution of an employee's name, the clause restricting the search to URLs that do not contain the domain of the organization separated organizational data from the rest of available information (one could also enlist all organizational domains, each in separate *-inurl* clause). As the second step of acquiring the evidence of a certain type, we send the query to a web search service and regard *the number of returned results* as a measure of personal expertness. Due to the limits of the Search Engine API technology we used, we had to restrict the number of persons for which we extracted global expertise evidence: it was unrealistic and unnecessary to issue thousands of queries containing each person for each query provided by a user. So, making an initial expert finding run on enterprise data was a requirement. As a result of that run, we used 100 most promising candidate experts (actually, the maximum number of candidates per query allowed for a single TREC submission) for the further analysis. Apart from the ranking built on fully indexed organizational data, we built rankings using 6 different sources of expertise evidence from the Global Web: Global Web Search, Regional Web Search, Document-specific Web search, News Search (all via Yahoo! Web search API), Blogs Search and Books Search (via Google Blog and Book Search APIs). Our experiments demonstrated a substantial increase in performance when we used combinations of up to three rankings. The best combination was comprised of the Enterprise, Global Web and News based rankings.

Despite that the main idea was to combine various rankings, we used obviously naive measure of expertness. That is why in the present work, we focus on combination of only two rankings, Enterprise and Global Web based, but use various measures of quality of web results returned by Yahoo Global Web Search API in response to the above described queries. Some of statistics per URL (the domain size and the number of inlinks) are still extracted by means of Google Web Search API, since it has no limit on the number of queries per user IP and it was a decisive factor to complete our experiments in time.

3. MEASURING THE QUALITY OF A WEB SEARCH RESULT

After all the majority of expert finding approaches is based on measuring the quality of a person-specific result set returned by the search engine in response to a query. Person-specific means that it contains only those documents that have at least one mention of the certain person and its quality may be represented by various features: the number of

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
University of Twente at the TREC 2008 Enterprise Track: Using the Global Web as an expertise evidence source	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Twente,PO Box 217, 7500 AE,Enschede, The Netherlands,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**4**	

documents it contains or the sum of their relevance probabilities. A result set returned by a typical web search engine consists of a list of result items described by their URLs, titles and summaries (snippets). To measure the overall quality of a web search result, we should aggregate calculated quality measures for all or top-k result items:

$$Expertise(e) = \sum_{Item \in WebResultSet} Quality(Item) \quad (1)$$

Certainly, downloading web pages using URLs of web result items for the deeper analysis of web result quality may lead to the better perfomance, but in our experiments we restrict ourselves to quality measures calcualted just from the search result pages or using such page statistics that can be quickly acquired from a search engine without downloading the full content of a page. All measures that we considered in this paper could be classified into two types: query-dependent and query-independent.

3.1 Query independent quality measures

In our experiments we focused on four kinds of query independent quality measures of a result item (web page).

3.1.1 URL length

Previous studies indicated that URL length is inversely proportional to the usefulness of the page it refers to [5, 3]. We apply simple quality measure based on this assumption: $Quality(Item) = 1/\sqrt{Length(Item_{URL})}$. The URL length is expressed in levels: the number of backslashes in the URL after its domain part. It should be mentioned that expressing the URL length in symbols performed much worse in our preliminary experiments.

3.1.2 Inlinks for domain

Another quality estimate we used is an approximation of the result item's authority. Since it was impossible to calculate sophisticated web graph centrality measures and since pages themselves are not often linked by pages outside of their domain, we used a simple inlink authority measure for the domain of the result item, considering that in many other authority measures (e.g. Pagerank) this value anyway propagates to all pages hosted at the result item's domain: $Quality(Item) = Inlinks(Domain(Item))$. The authority estimate was acquired using the ***link:*** clause plus the domain name to query Google Web Search API that returned the number of pages citing the given domain.

3.1.3 Domain size

We also supposed that the importance of the domain which hosts the returned result page should also be expressed by its size: $Quality(Item) = \sqrt{Size(Domain(Item))}$ The main intuition was that large domains usually become so only due to the time and money spent on their maintenance what in turn demonstrates their respectability. The size estimate was acquired using ***site:*** clause plus the domain name to query Google Web Search API that returned the number of pages indexed by Google at the given domain.

3.1.4 Freshness

We supposed that a page's last date of modification shows how much trust we should put in expertise evidence found in it. Obviously, the freshness of expertise evidence implicitly indicates the freshness of candidate's expert knowledge. In our preliminary experiments it appeared that considering only those results that where at least once modified (or created) after 2006 was better than just treating all of them equally useful:

$$Quality(Item) = \begin{cases} 1, Year(Item) \geq 2006 \\ 0, Year(Item) < 2006 \end{cases}$$

3.2 Query dependent quality measures

The state-of-the-art methods, including one that we use to get Enterprise based ranking [2], often rank candidates by the sum of relevance probabilities of pages that contain their mentions. Since it is very time- and broadband-consuming to download all pages in the result list in order to measure their relevance, we use a very simple measure of an $Item$'s (URL, Title or Summary) relevance which we sum over the result list:

$$Quality(Item) = \frac{N(q, q \in Item \wedge q \in Q)}{N(q, q \in Q)} \quad (2)$$

what is the number of query terms q appearing in the result $Item$ divided by the number of terms in the query Q. Since it is hard to tokenize URLs, we just search for a query term as for a substring in this case.

4. RANK AGGREGATION

The problem of rank aggregation is well known in research on metasearch [6]. Since our task may be viewed as *people metasearch*, we adopt solutions from that area. In our previous experiments with different rank aggregation methods we found that the simplest approach is also the best performing [8]. To get the final score we just summed the negatives of ranks for a person from each source to sort them in descending order:

$$Expertise(e) = \sum_{i=1}^{K} -Rank_i(e) \quad (3)$$

This approach is often referred as Borda count [1]. In our previous work we just sorted all candidates' expertise estimates for each evidence source to get their source-specific ranks. In this work we assigned these ranks more smoothly. First, we considered that all candidates with zero expertise estimates are always assigned with the lowest negative rank possible in the system (-100 in our experiments, since we always start by taking top-100 candidates from the Enterprise based ranking). Second, we assigned equal ranks to the candidates with equal expertise estimates, since before they were given arbitrary ranks by the sorting algorithm.

5. EXPERIMENTS

The CERC collection was indexed by Lucene retrieval engine using Snowball stemmer at the text parsing stage. For the purpose of finding candidate experts, we extracted all email addresses from the collection with *csiro.au* domain and *firstname.lastname*-like first part. We also had a list of email addresses to be banned which were not personal, but organizational addresses (e.g. publishing.photos@csiro.au). After all, we had 3500 candidate experts in total. Later, in order to find an association between a candidate and a document, we searched for the candidate's full email address or

Ranking	MAP	MRR	P@5
Enterprise	0.362	0.508	0.220
Enterprise +			
WebNumOfResults	0.485	0.627	0.256
WebURLLenInLevels	0.386	0.532	0.216
WebInlinksForDomain	0.477	0.632	0.252
WebSizeForDomain	0.477	0.604	0.248
WebAfter2006	0.491	0.620	0.256
WebRelevURL	0.501	0.650	0.26
WebRelevTitle	0.488	0.634	0.26
WebRelevSummary	0.485	0.627	0.252

Table 1: The performance of TREC 2007 queries

full name in the document's text. For each TREC title query we retrieved 50 documents (using a language model based retrieval model [2]) that contained at least one candidate expert mentioned. Then we analyzed these documents with the state-of-the-art Enterprise based expert finding method (Balog's candidate-centric Model 2 described in [2] and used in our previous experiments with web expertise evidence [8]). Finally, we considered only top 100 candidates from the Enterprise based ranking and built Web based rankings only for those.

The results analysis is based on calculating popular IR performance measures also used in official TREC evaluations: Mean Average Precision (MAP), precision at top 5 ranked candidate experts (P@5) and Mean Reciprocal Rank (MRR). We analyzed the performance of the Enterprise based ranking combined with one of the following rankings:

- **WebNumOfResults**: based on the number of web result items returned,
- **WebURLLenInLevels**: based on the sum of URL Length based quality estimates for web result items,
- **WebInlinksForDomain**: based on the sum of inlinks of domains of web result items,
- **WebSizeForDomain**: based on the sum of sizes of domains of web result items,
- **WebAfter2006**: based on the number of web result items modified or created after 2006,
- **WebRelevURL**: based on the sum of URL relevance probabilities for web result items,
- **WebRelevTitle**: based on the sum of title relevance probabilities for web result items,
- **WebRelevSummary**: based on the sum of summary relevance probabilities for web result items,

Our initial intention was to improve the combination of the **Enterprise** and the **Enterprise+WebNumOfResults** rankings that we regarded as our baseline (see Table 1). Only the **WebURLLenInLevels** ranking showed significantly degraded performance, the others were equally or better performing. Three rankings appeared to have slightly better performance in combination with the **Enterprise** rankings: **WebAfter2006**, **WebRelevTitle**, **WebRelevURL**. We also tried to further combine different rankings from the above list. However, we did not succeed to beat the **WebRelevURL**'s ranking performance with any of these combinations.

Ranking	MAP	MRR	P@5
Enterprise +			
WebNumOfResults	0.371	0.740	0.469
WebAfter2006	0.370	0.743	0.458
WebRelevURL	0.373	0.765	0.487
WebRelevTitle	0.371	0.754	0.480

Table 2: The performance of TREC 2008 queries

We finally submitted combinations of the **Enterprise** ranking with **WebNumOfResults**, **WebAfter2006**, **WebRelevTitle**, and **WebRelevURL** rankings as runs to TREC 2008 (see Table 2). The only difference with experiments with TREC 2007 queries is that we used our own infinite random walk based expert finding method [9] to build the **Enterprise** ranking. In this case all methods were equally effective according to MAP measure, but according to MRR and P@5 measures, considering relevance of URLs was indeed beneficial.

6. RELATED WORK

The usefullness of query-independent document quality measures for expert finding was recently studied. MacDonald et. al. [7] reported a bit different findings for the enterprise data only (e.g. all inlinks are only from pages of the same domain): they used similar expert finding method as a baseline and using Inlinks and URL length improved its MAP by a few percents. Similar document quality measures for document retrieval task can be found in some groups' reports on TREC Enterprise Track 2007 [12, 4, 11]. Measuring the quality of web result set to predict users' satisfaction with a search engine was just proposed by White et. al. [10].

7. CONCLUSIONS

The presented study demonstrates the predicting potential of the expertise evidence that can be found outside of the organization. We discovered that combining the ranking built solely on the Enterprise data with the Global Web based ranking may produce significant increases in performance. However, our main goal was to explore whether this result can be further improved by using various quality measures to distinguish among web result items. While, indeed, it was beneficial to use some of these measures, especially those measuring relevance of URL strings and titles, it stayed unclear whether they are decisively important.

There still stays a number of parallel directions to follow. First, various normalization and smoothing techniques could be applied to the URL quality measures we used. However, it seems more promising to apply machine learning mechanisms to find out which quality features of a web result item are the most important and how to combine them into a powerful expertise prediction model. Other sources of web expertise evidence besides Global Web should also not be overlooked: blog features (e.g. number of subscribers) when using Blog search based evidence or publication features (e.g. publisher's authority or a citation index) when using academic search services.

8. ACKNOWLEDGEMENTS

We want to sincerely thank Henning Rode for the help with the collection preprocessing and series of exciting discussions.

9. REFERENCES

[1] J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–284, New York, NY, USA, 2001. ACM.

[2] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06*, pages 43–50, 2006.

[3] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, New York, NY, USA, 2005. ACM.

[4] H. Duan, Q. Zhou, Z. Lu, O. Jin, S. Bao, Y. Cao, and Y. Yu. Research on enterprise track of trec 2007 at sjtu apex lab. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

[5] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, New York, NY, USA, 2002. ACM.

[6] Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li. Supervised rank aggregation. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 481–490, New York, NY, USA, 2007. ACM.

[7] C. Macdonald, D. Hannah., and I. Ounis. High quality expertise evidence for expert search. In *Proceedings of 30th European Conference on Information Retrieval (ECIR08)*, 2008.

[8] P. Serdyukov and D. Hiemstra. Being omnipresent to be almighty: The importance of the global web evidence for organizational expert finding. In *FCHER'08: Proceedings of the SIGIR'08 Workshop on Future Challenges in Expertise Retrieval*, 2008.

[9] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *CIKM '08*, Napa Valley, USA, 2008.

[10] R. W. White, M. Richardson, M. Bilenko, and A. P. Heath. Enhancing web search by promoting multiple search engine use. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, New York, NY, USA, 2008. ACM.

[11] M. Wu, F. Scholer, M. Shokouhi, S. Puglisi, and H. Ali. Rmit university at the trec 2007 enterprise track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

[12] J. Zhu, D. Song, and S. Rger. The open university at trec 2007 enterprise track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.