Being Omnipresent To Be Almighty: The Importance of the Global Web Evidence for Organizational Expert Finding

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

ABSTRACT

Modern expert finding algorithms are developed under the assumption that all possible expertise evidence for a person is concentrated in a company that currently employs the person. The evidence that can be acquired outside of an enterprise is traditionally unnoticed. At the same time, the Web is full of personal information which is sufficiently detailed to judge about a person's skills and knowledge. In this work, we review various sources of expertise evidence outside of an organization and experiment with rankings built on the data acquired from six different sources, accessible through APIs of two major web search engines. We show that these rankings and their combinations are often more realistic and of higher quality than rankings built on organizational data only.

Categories and Subject Descriptors:
H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval.

General Terms:
Algorithms, Measurement, Performance, Experimentation.

Keywords:
Enterprise search, expert finding, web search, blog search, news search, academic search, rank aggregation.

1. INTRODUCTION

In large organizations users often search for personalities rather than for relevant documents. In cases when required information is not published or protected, asking people becomes the only way to find an answer [14]. Experts are always in demand not only for short inquiries, but also for assigning them to some role or a job. Conference organizers may search for reviewers, company recruiters for talented employees, even consultants for other consultants to redirect questions and not lose clients [28].

The need for well-informed persons is often urgent, but the manual expert identification through browsing documents or via informal social connections is hardly feasible for large and/or geographically distributed enterprises. A standard text search engine cannot perform this task effectively. Instead, an *expert finding system* assists in the search for individuals or departments that possess certain knowledge and skills within the enterprise and outside [37]. It allows to save time and money on hiring a consultant when a company's own human resources are sufficient. Similarly to a typical search engine, an automatic expert finder uses a short user query as an input and returns a list of persons sorted by their level of knowledge on the query topic.

Expert finding started to gain its popularity at the end of '90s, when Microsoft, Hewlett-Packard and NASA published their experiences in building such systems [15, 16, 9]. They basically represented repositories of skill descriptions of their employees with simple search functionality. Nowadays these and other companies invest a lot to make their expert search engines commercially available and attractive [1, 2, 20]. Some large-scale free on-line people search[1] and expert finding[2] systems are already quite well-known in consultancy business [20]. On-line resume databases[3] and prominent social networks [4] are also often used to find professionals.

Apart from causing the new boom on the growing enterprise search systems market, expert finding systems also compelled close attention of the IR research community. The expert search task was introduced as a part of the Enterprise track of the Text REtrieval Conference (TREC) in 2005 [13]. Since that time, expert finding research blossomed, being conducted on the Enterprise TREC data in almost all cases. However, despite that a lot of research was produced outside of the TREC conference, the evidence of personal expertness mined from the TREC data was never combined with evidences acquired from other sources.

While Intranet of an organization still should be seen as a primary source of expertise evidence for its employees, the amount and quality of supporting organizational documentation is often not sufficient. At the same time, leading people search engines, such as `Zoominfo.com` or `wink.com` claim that none of their information is anything that one couldn't find on the Web [4]. Neglecting expertise evidence which can be easily found within striking distance is not practical.

In this study we propose to overcome the above-mentioned shortcomings and explore the predicting potential of expertise evidence acquired from sources publicly available on the Global Web. Using APIs of two major web search engines, we show how different types of expertise evidences, found

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

[1] `www.spock.com`
[2] `www.zoominfo.com`
[3] `www.monster.com`
[4] `www.linkedin.com`

in an organization and outside, can be extracted and combined together. Finally, we demonstrate how taking the web factor seriously significantly improves the performance of expert finding in an enterprise.

The remainder of this paper is organized as follows. The related research on expert finding is described in detail in the next section. In Section 3 we explain our strategy of expertise evidence acquisition from the Global Web. In Section 4 we show how we combine evidences from different web sources. Section 5 presents our experiments, Section 6 raises a discussion about our experimental results and expectations for the future, Section 7 outlines main conclusions and directions for the follow-up research.

2. NATURE OF EXPERTISE EVIDENCE

Finding an expert is a challenging task, because expertise is a loosely defined and not a formalized notion. It is often unclear what amount of personal knowledge may be considered enough to name somebody "an expert". It depends not only on the specificity of the user query, but also on characteristics of respective expertise area: on its age, depth and complexity. It is observed that on average people need at least ten years of experience to be experts in a given field [11]. The relevance of documents related to a person usually becomes the main evidence of the personal expertise. However, since the relevance can be determined only with some uncertainty, the expertise of a person appears to be even more uncertain. Even related content is not always a reliable evidence, since it may, for instance, contain discussions, showing the interest of involved people, but not their competence.

However, it is common to consider that the more often a person is related to the documents containing many words describing the topic, the more likely we may rely on this person as on an expert. The proof of the relation between a person and a document can be an authorship (e.g. we may consider external publications, descriptions of personal projects, sent emails or answers in message boards), or just the occurrence of personal identifiers (names, email addresses etc.) in the text of a document. Thus, the most successful approaches to expert finding obtain their estimator of personal expertise by summing the relevance scores of documents directly related to a person [36, 6]. In some works, only the score of the text window surrounding the person's mentioning is calculated [39]. In fact, these methods can be regarded as graph based since they measure personal expertness as a weighted indegree centrality in a topic-specific graph of persons and documents as it was previously done on a document-only network [25]. Some authors actually experimented with finding experts by calculating centralities in the person-only social networks [12, 44].

3. ACQUIRING EXPERTISE EVIDENCE FROM THE GLOBAL WEB

The main goal of this study is to answer the following research questions. First, what sources of expertise evidence outside of an organization are available? In what way should they be accessed? How to extract the expertise evidence from each source? What measures can be used to estimate expertness from the Global Web? Second, are these sources useful for finding experts? Is there any benefit in combining organizational and global expertise evidences?

The organization that we used for the study was CSIRO, Australia's Commonwealth Scientific and Industrial Research Organization. It has over 6 000 staff spread across 56 Australian sites and overseas. We used only publicly available documents - the crawl of *csiro.au* domain as it was provided by the Enterprise TREC community (see the detailed description of the data in Section 5).

3.1 Finding expertise evidence on the Web

The obvious solution for finding expertise evidence outside of the enterprise is to search for it in Global Web. There are basically two ways of doing that.

Crawling and RSS Monitoring. Many web data mining systems rely on focused crawling and analyzing discovered RSS feeds [23, 45]. It is often not even necessary to develop own web spider - topical monitoring can be implemented by means of such powerful aggregating tools as Yahoo! Pipes[5] or Google Alerts[6].

Search Engine APIs. Another much more comfortable way to "download the Internet" is to use open APIs of the famous web search engines - Google[7], Yahoo[8] or Live Search[9] [38]. Google has no limits on number of queries/day, Yahoo limits it to 5000, Live Search to 25000. All engines provide the access not only to their basic web search services, but also to search in *maps*, *images*, *news* etc. Unfortunately, it is not possible to automate data collection from services not accessible via APIs, even when it is easy to create wrappers for their web interfaces. Search engines usually have a right to ban IPs sending automated queries according to their Terms of Service.

3.2 Our evidence acquisition strategy

Since it is basically infeasible even for a wealthy organization to maintain an effective web search crawler, we focus on using APIs of two leading web search engines: Yahoo! and Google (Live Search API is still in unstable beta state). We extract expertise evidence for each person from their databases using the following strategy.

First, we build a query 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 (in case of Web or News search): *-inurl:csiro.au.*

Adding the organization's name is important for the resolution of an employee's name: the ambiguity of personal names in web queries is a sore subject. It was shown that adding the personal context to the query containing a name or finding such context automatically significantly improves the retrieval performance [40]. Of course, one could easily improve by listing names of all organizations where the person was ever employed (using OR clause) or by adding such context as the person's profession or title. However, the

[5] `pipes.yahoo.com`
[6] `google.com/alerts`
[7] `code.google.com/apis`
[8] `developer.yahoo.com/search/web/`
[9] `dev.live.com/livesearch/`

latter may still decrease the recall, cause this information is rarely mentioned in informal texts. It is also possible to apply more sophisticated strategies for names representation (e.g. using first name's diminutive forms and abbreviations), but we avoided using them for the sake of fast implementation and also as a quick solution for ambiguity resolution. In some cases, namely when using Global Web and News search services, we also added a clause restricting the search to URLs that do not contain the domain of the organization. It was done to separate organizational data from the rest of available information. In some cases, when an organization's domain is not unique, it is useful to just 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 one of the web search services, described further in this section. The returned *number of results* is considered as a measure of personal expertness. In other words, we ask a specific search engine: "Please, tell us how many times *this person* occurs in documents containing *these query terms* and not hosted at *the domain of her/his own organization*". The answer shows the degree of relation of a person to the documents on the topic what is a common indicator of personal expertness (see Section 2). Our technique is akin to the Votes method measuring a candidate's expertness by the number of organizational documents retrieved in response to a query and related to the candidate [36].

Due to limits of the Search Engine API technology we used, we had to restrict the number of persons for which we extracted global expertise evidence. In case of CSIRO, 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 from 20 to 100 most promising candidate experts per query for the further analysis. Processing one query takes less than a second. So, it usually took from 15 to 70 seconds to issue queries for all candidates, to wait for all responses of one search engine and to download all search result pages.

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). We describe each type of evidence and details of its acquisition further in this section.

3.3 Acquiring evidence from Enterprise

Despite the presence of vast amount of personal web data hosted outside of the corporate domain, the enterprise itself stays the main repository of structured and unstructured knowledge about its employees. Moreover, large part of enterprise documentation is often not publicly accessible and hence not indexed by any of web search engines. Even traditionally public Web 2.0 activities are often insistently popularized to be used fully internally within organizations for improving intra-organizational communication [24]. According to recent surveys [32], 24% of companies have already adopted Web 2.0 applications. Internal corporate blogging [27] and Project Wiki technologies [10] are the most demanded among them. For instance, it is reported that Microsoft employees write more than 2800 blogs and about 800 of them are only internally accessible [18].

Since it is usually possible to have fast access to the content of indexed documents in an Enterprise search system, we build an Enterprise search based ranking using state-of-the-art expert finding algorithm proposed by Balog et. al. [6]. It measures candidate's expertness by calculating a weighted sum of scores of documents retrieved to a query and related to the candidate:

$$Expertise(e) \approx \sum_{D \in Top} P(Q|D)P(e|D) \tag{1}$$

$$P(e|D) = \frac{a(e,D)}{\sum_{e'} a(e',D)}, \tag{2}$$

where $P(Q|D)$ is the probability that the document D generates the query Q, measuring the document relevance according to the probabilistic language modeling principle of IR [26], $P(e|D)$ is the probability of association between the candidate e and the document D, $a(e,D)$ is the non-normalized association score between the candidate and the document proportional to their strength of relation. Note that the difference with the measure we use to aggregate expertise evidence from the Global Web (simple count of all documents matched to a query and related to the person) is that we consider all document scores equal. We also do not assume that the amount of that document score propagated to a mentioned candidate depends on the number of candidates in a document. The described ranking method represents a baseline in our experiments.

3.4 Acquiring evidence from Web search

The importance of the Global Web for finding information about people is unquestionable. Especially, since people recently started to care much about their "online reputation"[10]. Everyone wants to be found nowadays and it is often crucial to be searchable in the Internet Era. The word "Google" is officially added to the Oxford English Dictionary as a verb. "Googling" a person is one of the most popular search activities with dedicated manuals and howtos [41]. 30% of all searches on Google or Yahoo! are for specific people or people related [4]. The increasingly used practice for employment prescreening is to "Google" applicants [29]. A 2006 survey conducted by `CareerBuilder.com` found that one in four employers use Internet searches to learn more about their potential employees and actually more than half of managers have chosen not to hire an applicant after studying their online activity.

There is however a huge controversy on what search engine is better: Google or Yahoo! Almost everyone has his own opinion on this topic. From one point of view, Google has much larger share of searches in U.S. (59% in February 2008[11]), but Yahoo! is still a bit ahead of Google according to The American Customer Satisfaction Index[12]. To avoid following the common path, we preferred Yahoo! Web Search API over Google. The reason was also that Yahoo's search APIs are more developer-friendly and, although they have some usage limitations (see Section 3.1), they offer more features and they are more flexible, by also including XML output.

[10] `www.manageyourbuzz.com/reputation-management/`

[11] `www.comscore.com`

[12] `www.theacsi.org`

In order to analyze different scopes of a person's mentioning on the web, we built expertise rankings based on several kinds of web searches: without any restrictions (except those mentioned in Section 3.2) and with restrictions on domains location and on the type of documents:

- **Global Web Search**. The search without restriction of the scope.
- **Regional Web Search**. The search only at web-sites hosted in Australia (by using Yahoo's *country* search option). The purpose was to study whether we may benefit by expanding the search scope gradually, first searching for the expertise evidence in a company's region.
- **Document-specific Web Search**. The search only in PDF documents (by using Yahoo's *format* search option). The purpose was to study whether it is beneficial to differentiate document types. The PDF format was selected as a de-facto standard for official on-line documents (white papers, articles, technical reports) that we regarded as one of the main sources of expertise evidence.

3.5 Acquiring evidence from News Search

Good experts should be a bit familiar to everybody. However, to be searchable and broadly represented on the Web does not always mean to be famous and authoritative. What really matters is to be on everyone's lips, to be on the top of the news. First, it is well-known that news reflect internet buzzes, especially in blogosphere, serving as a filter for events and topics interesting for a broad audience (and vice versa is also true) [34]. Second, being on the news often means to be distinguished for your professional achievements: for making a discovery, starting a trend, receiving an award.

Yahoo![13], Google[14] and Live Search offer APIs for their News Search services. However, their significant limitation making them useless for expertise evidence acquisition is that they allow to search only in news from the past month. Since employees are not celebrities and hence are not mentioned in news daily, it is almost impossible to extract sufficient expertise evidence from these services. Google also has News Archive Search[15], but has no API for accessing it.

To realistically simulate the usage of News Search, we took our usual query (see Section 3.2), added *inurl:news* clause to it and sent it to Yahoo! Web Search service. In this way we restricted our search to domains and sub-domains hosting only news or to pages most probably containing only news.

3.6 Acquiring evidence from Blog Search

As it was already mentioned in Section 3.3, blogs are very rich sources of knowledge about personal expertise. The larger part of corporate professional blogs is public and indexed by major blog search engines. Leading recruiting agencies predict the rapid increase of interest in candidates passionate about writing their blogs [22]. Actually, the retrieval task of finding relevant blogs quite resembles the task of finding experts among bloggers in the Blogosphere. Recently, Balog et. al. successfully experimented with expert finding methods for *blog distillation* task on TREC 2007 Blog track data [7].

Two major blog search engines are fiercely competing with each other leaving others far behind: Technorati and Google Blog Search. According to the spreading Internet hype and recent random probings Google has significantly better coverage for blogs [42]. Its Blog Search API is much more developer-friendly than Technorati's, which is often reported to be very unreliable (and it was even impossible to get an Application ID at `technorati.com/developers` at the time of writing this paper). Despite that Google Blog Search API also has its own inconvenient limitations (it can only return up to 8 links in result set), we use it for building Blog Search based ranking (see Section 3.2).

3.7 Acquiring evidence from Academic Search

Academic publications is a great source of expertise evidence, especially for R&D companies such as CSIRO. Not all of them can be found at corporate web-sites, since their public distribution may be forbidden by copyright terms. There are two major multidisciplinary Academic Search engines: Google Scholar [16] and Live Search Academic[17]. The others like *Scopus* or *Web of Science* index significantly less publications on many subjects, do not consider unofficial publications and are sometimes restricted to specific types of articles (e.g. to journals). Several studies have shown that it is effective to calculate bibliometric measures for estimating reputation of scientists using citations found in Google Scholar [8]. It also becomes more popular among researchers to specify in their resumes the number of citations in Google Scholar for their publications. Google Scholar can actually be regarded as a ready-to-use expert finding system, since it always shows 5 key authors for the topic at the bottom of the result page.

Unfortunately, there is no possibility to access any academic search engine via API. However, Google provides API for a very similar search service: Book Search[18]. While its publication coverage is not as large as Google Scholar's, there is a high overlap in the data they both index, since Google Scholar always returns items indexed by Book Search for non-fiction subjects. Using Books Search also naturally allows to search for expertise evidence in not strictly academic sources. So, we build an Academic Search based ranking by sending queries (see Section 3.2) to Google Book Search service.

4. COMBINING EXPERTISE EVIDENCES THROUGH RANK AGGREGATION

The problem of rank aggregation is well known in research on metasearch [33]. Since our task may be viewed as *people metasearch*, we adopt solutions from that area. We also decided to use only ranks and ignore the actual number of results acquired for each candidate expert and a query from each search service. It was done for the sake of comparability and to avoid the need for normalization of values.

In our preliminary experiments with different rank aggregation methods we found that the simplest approach is also the best performing. To get the final score we just sum the

[13] `news.yahoo.com`

[14] `news.google.com`

[15] `news.google.com/archivesearch`

[16] `scholar.google.com`

[17] `academic.live.com`

[18] `books.google.com`

	Baseline	YahooWeb	YahooWebAU	YahooWebPDF	YahooNews	GoogleBlogs
YahooWeb	0.287					
YahooWebAU	0.254	0.502				
YahooWebPDF	0.259	0.513	0.359			
YahooNews	0.189	0.438	0.400	0.395		
GoogleBlogs	0.069	0.424	0.412	0.422	0.494	
GoogleBooks	0.111	0.419	0.411	0.412	0.453	0.202

Table 1: The normalized Kendall tau distance between all pairs of rankings

negatives of ranks for a person from each source to sort them in descending order:

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

This approach is often referred as Borda count [5]. We also tried to learn weights of sources with the Ranking SVM algorithm, using its SVM^{map} version which directly optimizes Mean Average Precision[19] [43]. However, its performance was surprisingly nearly the same as Borda count's.

5. EXPERIMENTS

We experiment with the **CERC** collection used by the Enterprise TREC community in 2007. It represents a crawl from Australia's national science agency's (CSIRO) web site. It includes about 370 000 web documents (4 GB) of various types: personal home pages, announcements of books and presentations, press releases, publications. Instead of a list of candidate experts, only the structure of candidates' email addresses was provided: *firstname.lastname@csiro.au*. Using this as a pattern we built our own candidates list by finding about 3500 candidates in the collection. 50 queries with judgments created by CSIRO Science Communicators (a group of expert finders on demand) were used for the evaluation. At the collection preparation stage, we extract associations between candidate experts and documents. We use simple recognition by searching for candidates email addresses and full names in the text of documents. For the CSIRO documents the association scores $a(e, D)$ between documents and found candidates are set uniformly to 1.0 (see Section 3.3).

The results analysis is based on calculating popular IR performance measures also used in official TREC evaluations: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and precision at top 5 ranked candidate experts (P@5). MAP shows the overall ability of a system to distinguish between experts and non-experts. P@5 is considered more significant than precisions at lower ranks since the cost of an incorrect expert detection is very high in an enterprise: the contact with a wrong person may require a mass of time. If we consider that the user can be satisfied with only one expert on the topic (considering that all experts are always available for requests), then the performance of MRR measure becomes crucial.

In our experiments discussed below we compare our methods with a baseline ranking and also study the effectiveness of combinations of rankings. The performance of the following rankings and their combinations is discussed further:

- **Baseline**: Baseline Enterprise search based ranking (see Section 3.3),
- **YahooWeb**: Yahoo! Global Web search based ranking (see Section 3.4),
- **YahooWebAU**: Yahoo! Regional Web search based ranking (see Section 3.4),
- **YahooWebPDF**: Yahoo! Document-specific Web search based ranking (see Section 3.4),
- **YahooNews**: Yahoo! News search based ranking (see Section 3.5),
- **GoogleBlogs**: Google Blog search based ranking (see Section 3.6),
- **GoogleBooks**: Google Book search based ranking (see Section 3.7).

	MAP	MRR	P@5
Baseline	0.361	0.508	0.220
YahooWeb	**0.423**	0.547	**0.248**
YahooWebAU	0.372	0.462	0.220
YahooWebPDF	0.358	0.503	0.200
YahooNews	0.404	0.554	0.216
GoogleBlogs	0.406	**0.582**	0.200
GoogleBooks	0.373	0.517	0.200

Table 2: The performance of rankings

Before starting analyzing the quality of each ranking, we compare them using normalized Kendall tau rank distance measure [19]. As we see in Table 1, the **Baseline** ranking appears to be very similar to the **GoogleBlogs** and **GoogleBooks** rankings. While the similarity of the latter is also supported by its similar performance with the **Baseline** (see Table 2), the **GoogleBlogs** obviously improves the **Baseline** not being considerably different. It probably happens because it is different mostly at more important lower ranks. It is also interesting that all four rankings acquired using the same Yahoo Web Search API differ very substantially. This result approves that at least the decision to segregate different information units within one source was reasonable. On the contrary, rankings acquired from Google and even from its different search services disagree at a much lower level. We may suppose that it is explained by the fact that both sources provide only a limited amount of evidence. The Google Blog Search API returns at maximum 8 results, so all candidate experts mentioned more than 8 times in blogs are regarded equal. Google Book search basically allows us to distinguish only between noted specialists and does not provide us with all sorts of academic expertise evidence.

The performance of each ranking is presented in Table 2. We see that restricting the scope of web search to the regional web or to specific file format does not lead to better results. Both the **YahooWebAU** and the **YahooWebPDF** rankings are inferior to the **YahooWeb** ranking and to the

[19] `projects.yisongyue.com/svmmap/`

YahooWeb +	MAP	MRR	P@5
Baseline	**0.460**	**0.604**	0.240
YahooWebAU	0.390	0.483	0.224
YahooWebPDF	0.402	0.525	0.208
YahooNews	0.406	0.543	0.232
GoogleBlogs	0.427	0.562	0.223
GoogleBooks	0.452	0.567	**0.244**

Table 3: The performance of combinations of the YahooWeb ranking with the other rankings

	MAP	MRR	P@5
YahooWebPDF + GoogleBooks	0.440	0.567	0.232
YahooNews + GoogleBlogs	0.420	0.571	0.216

Table 4: The performance of additional combinations inferring better Academic and Social Media evidences

Baseline. However, all other rankings built on web evidence are better than the **Baseline** in terms of MAP and MRR measures. It is hard to decide which of them is the best: **YahooWeb** is much better in MAP and P@5, but if user needs to detect the most knowledgeable person fast, using evidence from news and blogs seems a better idea according to the performance of the MRR measure. The **GoogleBlogs** ranking outperforms the baseline only slightly, so its use without combining it with other evidences is questionable.

We also experimented with combinations of rankings (see Section 4). Following the principle that we should give a priority to the best rankings, we combined the most effective **YahooWeb** ranking with each other ranking (see Table 3). We surprisingly found that the combinations of that ranking with the **Baseline** and the **GoogleBooks** rankings, which are not the best alone, are the best performing. Probably, since according to the normalized Kendall tau distance (see Table 1) these rankings are more similar to the **YahooWeb** ranking, their combination produces a more consistent result. We also combined the **Baseline** ranking with each one another, but found that its combination with the **YahooWeb** ranking is still the best.

In order to study the future potential of web evidence combinations, we decided to simulate the inference of web evidences which we can not currently acquire through APIs. First, we combined the **YahooWebPDF** and the **GoogleBooks** rankings to infer a better academic search based evidence. Considering that a lot of official and unofficial publications are publicly accessible in PDF format, we hoped to simulate the output of Google Scholar-like search service. As we see in Table 4, the performance of that combined ranking approved our expectations: it is better than each of these rankings used alone. Second, we tested the combination of the **YahooNews** and the **GoogleBlogs** rankings considering that it would represent an output from some future Social Media search service as it is envisioned by many [21]. The advantage of this combination is visible, but less obvious. It is certainly better than the **YahooNews** ranking, but outperforms the **GoogleBlogs** ranking only according to the MAP measure.

As we see in Table 5, further combination showed that when we combine the **Baseline** ranking, the **YahooWeb** ranking and the **YahooNews** ranking, we get improvements for the MAP and the P@5 measures. In total using that combination we had 29% improvement of MAP, 20% of MRR, and 14% of P@5. Combinations of 4 and more rankings only degraded the performance. To test statistical significance of the obtained improvement, we calculated a paired t-test. Results indicated that the improvement is significant at the $p < 0.01$ level with respect to the baseline.

YahooWeb + Baseline +	MAP	MRR	P@5
YahooWebAU	0.463	**0.606**	0.240
YahooWebPDF	0.446	0.589	0.240
YahooNews	**0.468**	0.600	**0.252**
GoogleBlogs	0.452	0.591	0.244
GoogleBooks	0.449	0.597	0.232

Table 5: The performance of combinations of the YahooWeb and the Baseline rankings with the other rankings

6. DISCUSSION

As it was demonstrated by our experiments, we are able to gain significant improvements over the baseline expert finding approach which analyzes the data only in the scope of an organization. We found that the quality of inference of personal expertness is proportional to the amount of expertise evidence. When we search for this evidence also outside of an organization in the Global Web, we increase our potential to guess about competence of its employees. It was also clear from experiments that combining different sources of evidence through simple rank aggregation allows to improve even more. This improvement is also probably caused by diminishing of the ranks of persons that appear in organizational documentation accidentally or by technical and bureaucratic reasons (e.g. web-masters or secretaries). Such persons not actually related to the topic of a query seemingly are only locally frequent and do not appear often in each source. The results of our investigation suggest that it is promising to discover more sources of expertise evidences and to improve the quality of evidence acquired from these sources.

6.1 Finding new sources of expertise evidence

While we focused our studies on the predefined subset of search services selected by their popularity and supposed richness in expertise evidence, there are more sources. Some of them are already able to provide some expertise evidence, but for companies with different specialization than CSIRO's. Other ones are currently not as popular and all-embracing, but are on the rise of their authority.

Social Networks. Social networks is an indispensable source of knowledge about personal skills and experience. They allow to extract expertise evidence not solely from a user profile, but also from its context: directly "befriended" user profiles or profiles connected implicitly through sharing the same attributes (e.g. places of work or visited events). However, while such huge networks as `LinkedIn.com` (more than 17 million members) and `Facebook.com` (more than 70 million members) are very popular for finding specialists to recruit them [30, 31], it is still hard to compare employees within organization using this information since simply not all of them have their own account there.

Expert databases. Those who are not willing to create their own professional profile, will be supplied with one.

Such repositories of experts as Zoominfo.com and many others [4] automatically summarize all information about people found on the Web to make them searchable. Many of them provide APIs for programmatic access to their databases[20].

Vertical Search Engines. Specialized topic-oriented search engines should be helpful for finding experts in specific industries: SearchFinance.com - for finding economists, Medstory.com - for doctors, Yahoo! Tech[21] and Google Code Search[22] - for software engineers etc.

User generated content. There are other ways to share expertise besides blogging. Giving professional advice at Yahoo![23] or LinkedIn[24] Answers or authoring Wikipedia articles [17] are activities that indicate personal proficiency not only by their content, but also by feedback of involved users assessing the quality of advice [3]. There are also communities like Slideshare.com where knowledge exchange is accomplished with the minimum effort by just uploading personal presentation slides.

6.2 Improving the quality of evidence

In this work we used used a simple measure of personal expertness counting the number of information units in a source that contain all query terms and a candidate mention. Since we consider every link returned by a search service as a partial evidence of personal expertness, the next step would be to differentiate the strength of these evidences by taking various properties of these links into account.

Considering relevance of links. The state-of-the-art expert finding approaches go beyond simple counting of candidate's mentions in documents on a topic and sum relevance scores of all related documents (see Section 2). In our case it is hard to measure the relevance of returned links without downloading entire documents (what is not possible sometimes, e.g. for links to paid content). However, we can think about some options. We may try to measure relevance of web snippets returned together with links. It is possible to issue a query without a person's name within and get only topic based ranks of documents. But since most engines return only first thousand of matched pages, that strategy may fail for non-selective short ambiguous queries producing significantly larger result.

Considering authority of links. It was recently proposed to measure the strength of expertise evidence extracted from a web page by the number of its inlinks [35]. There are web services providing similar statistics: Yahoo! Search API (Site Explorer) returns the number of inlinks for a provided URL, sites like Prchecker.info even show the estimate of Google PageRank. Academic search engines like Google Scholar usually return the number of citations per publication in their result set.

Considering popularity of links. The click/visit popularity is also a primary evidence of web page quality. Not only major search engines with their huge query logs are able to analyze such statistics. Web sites like Alexa.com and Compete.com provide an unique opportunity (also through API) to inquire about a total number of visits and overall time spent at a domain by web surfers.

7. CONCLUSIONS AND FUTURE WORK

In this paper we proposed a way to gather additional expertise evidence apart from that available in the organization. We used various kinds of Global Web search services to acquire a proof of expertness for each person which was initially pre-selected by an expert finding algorithm using only organizational data. By means of APIs of two major search engines, Yahoo! and Google, we built six rankings of candidate experts per query and demonstrated that rankings from certain web sources of expertise evidence and their combinations are significantly better than the initial enterprise search based ranking.

In the future we would like to explore the usefulness of other sources of expertise evidence and to apply more sophisticated measures than just a simple number of related topical information units per person in a source. It is also clear that we need a more efficient strategy of evidence acquisition. Sending queries for each person and a query to every web search service is not practical, resource consuming and causes too much latency. The round-robin strategy used in this work may be improved by asking evidence for less promising persons from each next evidence source after rank aggregation at each step.

8. REFERENCES

[1] IBM Professional Marketplace matches consultants with clients. White paper. November 2006.

[2] Enterprise search from Microsoft: Empower people to find information and expertise. White paper. Microsoft, January 2007.

[3] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 665–674, New York, NY, USA, 2008. ACM.

[4] M. Arrington. War of the people search. www.techcrunch.com/2007/05/09/war-of-the-people-search/.

[5] 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.

[6] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *SIGIR '07*, pages 551–558, 2007.

[7] K. Balog, M. de Rijke, and W. Weerkamp. Bloggers as experts. In *SIGIR '08*, 2008.

[8] J. Bar-Ilan. Which h-index? - a comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271, February 2008.

[9] I. Becerra-Fernandez. Facilitating the online search of experts at NASA using expert seeker people-finder. In *PAKM'00, Third International Conference on Practical Aspects of Knowledge Management*, 2000.

[10] M. Buffa. Intranet wikis. In *Proceedings of Intraweb workshop, WWW'06*, 2006.

[11] N. Charness. *The Cambridge Handbook of Expertise and Expert Performance.* Cambridge University Press, 2006.

[20] www.programmableweb.com/apis/

[21] tech.yahoo.com

[22] codesearch.google.com

[23] answers.yahoo.com

[24] linkedin.com/answers

[12] H. Chen, H. Shen, J. Xiong, S. Tan, and X. Cheng. Social Network Structure behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

[13] N. Craswell, A. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *Proceedings of TREC-2005*, Gaithersburg, USA, 2005.

[14] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. Panoptic expert: Searching for experts not just for documents. In *Ausweb Poster Proceedings*, Queensland, Australia, 2001.

[15] T. Davenport. Knowledge Management at Microsoft. White paper. 1997.

[16] T. Davenport. Ten principles of knowledge management and four case studies. *Knowledge and Process Management*, 4(3), 1998.

[17] G. Demartini. Finding experts using Wikipedia. In A. V. Zhdanova, L. J. B. Nixon, M. Mochol, and J. G. Breslin, editors, *FEWS*, volume 290 of *CEUR Workshop Proceedings*, pages 33–41. CEUR-WS.org, 2007.

[18] L. Efimova and J. Grudin. Crossing boundaries: A case study of employee blogging. In *HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, page 86, Washington, DC, USA, 2007. IEEE Computer Society.

[19] R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160, 2003.

[20] L. Fields. 3 great databases for finding experts. *The Expert Advisor*, (3), March 2007.

[21] C. Firestone, P. Kelly, and R. Adler. Next-Generation Media: The Global Shift. Report, Aspen Institute, 2007.

[22] T. Golta. The 2008 recruiting landscape. Five recruiting gurus' 2008 predictions. White paper. January 2008.

[23] D. Gruhl, D. N. Meredith, J. H. Pieper, A. Cozzi, and S. Dill. The web beyond popularity: a really simple system for web scale rss. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 183–192, New York, NY, USA, 2006. ACM.

[24] GuideWireGroup. Blogging in the enterprise. White paper. January 2005.

[25] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM.

[26] D. Hiemstra. *Using Language Models for Information Retrieval.* Phd thesis, University of Twente, 2001.

[27] J. Huh, L. Jones, T. Erickson, W. A. Kellogg, R. K. E. Bellamy, and J. C. Thomas. Blogcentral: the role of internal blogs at work. In *CHI '07: CHI '07 extended abstracts on Human factors in computing systems*, pages 2447–2452, New York, NY, USA, 2007. ACM.

[28] M. Idinopulos and L. Kempler. Do you know who your experts are? *The McKinsey Quarterly*, (4), 2003.

[29] M. Jones, A. Schuckman, and K. Watson. *The Ethics of Pre-Employment Screening Through the Use of the Internet*, chapter 4. Ethica Publishing, 2007.

[30] R. King. Social networks: Execs use them too. *BusinessWeek*, September 2006.

[31] E. Kolek and D. Saunders. Online disclosure: An empirical examination of undergraduate Facebook profiles. *NASPA Journal*, 45(1), 2008.

[32] J. Levine. Business gets social: Corporate usage of Web 2.0 explodes. ChangeWave, January 2008.

[33] 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.

[34] L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *Proceedings of the AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, page 8, Stanford, 2006.

[35] 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.

[36] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06*, pages 387–396, 2006.

[37] M. T. Maybury. Expert finding systems. Technical Report MTR06B000040, MITRE Corporation, 2006.

[38] F. McCown and M. L. Nelson. Agreeing to disagree: search engines and their public interfaces. In *JCDL '07: Proceedings of the 7th ACM/IEEE joint conference on Digital libraries*, pages 309–318, New York, NY, USA, 2007. ACM.

[39] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *CIKM '07nt*, pages 731–740, 2007.

[40] D. Shen, T. Walkery, Z. Zhengy, Q. Yangz, and Y. Li. Personal name classification in web queries. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 149–158, New York, NY, USA, 2008. ACM.

[41] C. Sherman. *Google Power: Unleash the Full Potential of Google. The art of googling people*, chapter 12. Barnes and Noble, 2005.

[42] M. Thelwall and L. Hasler. Blog search engines. *Online Information Review*, 31(4):467–479, 2007.

[43] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278, New York, NY, USA, 2007. ACM.

[44] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW '07*, pages 221–230, 2007.

[45] C.-N. Ziegler and M. Skubacz. Towards automated reputation and brand monitoring on the web. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 1066–1072, Washington, DC, USA, 2006. IEEE Computer Society.