Future Challenges in Expertise Retrieval

Preface

At the TREC Enterprise Track in 2005 the need to study and understand expertise retrieval was recognized through the introduction of an Expert Finding task (as opposed to mere document retrieval). The task has generated a lot of interest in the IR community, and rapid progress has been made in terms of modeling, algorithms, and evaluation over the past 3 years. In fact, expertise retrieval has reached the point where it is appropriate to assess progress, bring people from different research communities together, and define a research agenda for the next years. This ACM SIGIR Workshop on Future Challenges in Expertise Retrieval (fCHER) aims to determine what we have accomplished and where we need to go from here in expertise retrieval.

The workshop schedule lets accommodate regular papers (up to 8 pages long) along with position papers (up to 4 pages long). The program committee accepted 8 papers (4 full and 4 position papers). Each paper was reviewed by at least three members of the program committee. In addition, the fCHER program also includes an invited talk by Arjen P. de Vries, from CWI (The Netherlands).

Finally, we would like to thank the ACM and SIGIR for hosting the workshop; the ILPS group of the University of Amsterdam for the computing resources to host the workshop web site; all the authors who submitted papers for the hard work that went into their submissions; the members of our program committee for the thorough reviews in such a short period of time; Arjen P. de Vries for agreeing on giving an invited talk.

Krisztian Balog and Yong Yu
Program Committee Co-Chairs

Table of Contents

fCHER Organization

Program Co-Chair: Krisztian Balog, *(University of Amsterdam, The Netherlands)*
Yong Yu, *(Apex Knowledge & Data Management Lab, China)*

Program Committee: Arjen de Vries, *(CWI, The Netherlands)*
Bettina Berendt, *(Katholieke Universiteit Leuven, Belgium)*
ChengXiang Zhai, *(University of Illinois at Urbana-Champaign, USA)*
Craig Macdonald, *(University of Glasgow, UK)*
David Hawking, *(CSIRO, Australia)*
Elaine Toms, *(Dalhousie University, Canada)*
Henry Kautz, *(University of Rochester, USA)*
Ian Soboroff, *(NIST, USA)*
Irma Becerra-Fernandez, *(Florida International University, USA)*
Jianhan Zhu, *(Open University, UK)*
Jimmy Huang, *(York University, Canada)*
Jun Wang, *(University College London, UK)*
Leif Azzopardi, *(University of Glasgow, UK)*
Loren Terveen, *(University of Minnesota, USA)*
Maarten de Rijke, *(University of Amsterdam, The Netherlands)*
Mark S. Ackerman, *(University of Michigan, USA)*
Peter Bailey, *(Microsoft, USA)*
Peter Mika, *(Yahoo! Research Barcelona, Spain)*
Yutaka Matsuo, *(AIST, Japan)*

Finding People and Documents, Using Web 2.0 Data

Einat Amitay
einat@il.ibm.com

David Carmel
carmel@il.ibm.com

Nadav Golbandi
nadav.golbandi@gmail.com

Nadav Har'El
nyh@il.ibm.com

Shila Ofek-Koifman
shila@il.ibm.com

Sivan Yogev
sivany@il.ibm.com

IBM Haifa Research Lab
Haifa 31905, Israel

ABSTRACT

Given a user's free-text query, search engines return a ranked list of documents that are likely to be helpful to the user. In this research, we propose a simple yet highly effective technique for also providing a ranked list of *related people* to every search. The list of people related to the query is calculated at search time using an enhanced *faceted search* engine, based on person-document relationships mined from several Web 2.0 applications (such as blogs and social bookmarks) in the intranet of a large enterprise.

Our hypothesis is that the *related people* we retrieve for a query are people who have special interest in the query's topic, and thus may be useful to the person making this query. We conducted a large user study with over 600 people to confirm this hypothesis.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Social Search, Faceted Search, Enterprise Search

1. INTRODUCTION

When they are in need of information, some people like to find a written document which explains what they want to know. Yet, other people prefer to find the right person — one who might know the answer to their question — and ask him or her for the specific information they need. Most people are somewhere between these extremes, preferring to find documents in some cases, and people to ask in other cases. Also, for some topics, only one of these information sources is available. This is why we believe that search engines should provide both types of results: Given a query, the search engine should provide both a ranked list of documents that might answer the user's query, and a ranked list of people that are interested in the query's topic, and might be able to help.

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

Often the data and techniques used to find relevant documents, and those used to find relevant people (as in expert search), were separate and unrelated. In this paper, we propose a *unified* method that finds both relevant documents and relevant people for every query.

As we shall see in section 2 below, the key to our technique is knowing for each document which person is *related* to it. An excellent source for both documents and document-person relationships are so-called Web 2.0 applications, such as blogs and social bookmarking systems. In addition to the actual documents, these applications can tell us who is related to each document, and in what way. For example, a person can be related to a blog entry as its *author* or as a *commenter*, and can be related to any page as a *bookmarker*.

We will then show how to use this *social information* — documents and document-person relationships — to determine which people are most relevant to a given query. We will use an enhanced faceted search engine to determine the (potentially large) set of relevant documents for this query, and then which people are most related to these documents.

In this work, we focused our attention on the case of *enterprise search*, i.e., search in the intranet of a large organization. Compared to the open Internet, people in the enterprise are easier to track (because they use the same user-id everywhere), and are more likely to be helpful to each other.

In section 3, we evaluate the validity of the "related people" results. The individual document-person relationships (author, commenter, bookmarker) indicate that the person is in some way *relevant* to the content of the document; So it is natural to hypothesize that the query-person relationship we derive from them also measures the person's relevance to the topic of the query. We checked this hypothesis using a large user study with over 600 participants.

2. THE SYSTEM

2.1 Social Information

Traditionally, building a Web site took a considerable amount of effort and expertise, so most users were relegated to the role of information consumers, not producers. Web 2.0 services, such as forums, wikis, collaborative bookmarking services, and many more, allow ordinary users to become information producers. In turn, this allows people to learn from the experiences and knowledge of their peers — something which is especially important in the intranet of a large international enterprise.

Web 2.0 sources not only provide a new wealth of infor-

mation, they also provide new types of information, which we call *social information*. The new types of information include user-supplied metadata for documents (bookmarks, tags, ratings, comments), relationships between people and documents (who wrote a document, who commented on it, who tagged it, and so on), and other relationships such as between people and people, or between documents and tags.

The goal of a *social search engine* is to use the social information to improve the user's search experience over regular full-text search. One way of improving search is to improve the relevance of document results [3, 16, 6]: Tags (and other forms of comments) supply more text that can be considered during search, and important documents can be recognized by the amount of user activity around them (such as the number of times they were bookmarked or commented on).

But using the social information, we want to do more than just return better documents. The literature [5, 17, 10, 13, 15, 14] proposes the idea of *multi-entity search*, where other entities besides documents can be used in queries or turn up in search results. In our case, we want *people* to be searchable entities in our system, exactly like documents: Related people will be returned for every query (in addition to the relevant documents), and people can also be used as query terms.

2.2 Related People

In the standard vector space model of IR, each document is represented as a normalized vector that measures the relevance of each term (word) to the document. The entire document collection is therefore represented as a relevance matrix D between documents and terms; D_{ij} is the relevance of the ith document to the jth term. A query is represented, just like a document, as a vector q. The product Dq is a vector giving the relevance of each document to the query q. I.e., these are the search results.

The people-document relationships in the social information allow us to define a second relevance matrix P, between documents and people. P_{ij} measures the relevance of the ith document to the jth person. We might, for example, want to give a high relevance P_{ij} when person j wrote document i, a lower relevance if they commented it, and a lower still relevance if they merely bookmarked it.

Multiplying the term-document relevance matrix and the document-person relevance matrix yields a term-person relevance matrix P^TD that can be then be used in search: P^TD can be, just like D, multiplied by a query vector, resulting this time in relevant people (instead of documents). The relevance of these people to the query is *indirect*, through the documents — a person is relevant to a query if he or she are relevant to documents which are relevant to the query.

The need to calculate the matrix product P^TD causes problems, though. Every change to the social information can require modifying large parts of this matrix, making it difficult to index dynamically-changing data. It also means that searching for relevant documents and relevant people is done separately, using two different relevance matrices (D and P^TD). Finally, most search engines offer capabilities beyond the simple vector space model (e.g., supporting searches of multi-word phrases, considering term proximity, and more), and using the matrix P^TD directly forces us to give up on these features.

We therefore propose an alternative technique which (as we shall show) gives equivalent results, but solves all the above problems. The idea, already found in [2, 11], is to first use the given search engine to find the relevant documents; Then, knowing which people are relevant to each of these documents, we start aggregating the relevance of each person. This process can be realized using *faceted search*, with the related people added as a *facet* to each document:

2.3 Faceted Search for Related People

Faceted search is a commonly-used technique for adding navigation to a search engine. A *facet* is a single attribute of the document, e.g., in a book search application there might be an "Author" facet and a "Price" facet, and in our application there is a "Related Person" facet. Faceted search starts, like ordinary search, by finding all documents matching the user's query. But while an ordinary search system will only show the few documents with the highest relevance, a faceted search system goes over all matching documents, counting the number of documents found for each subcategory of the facet (individual authors, price ranges, etc.), and finally displays the categories with the highest counts.

Our unified search solution is based on a faceted search library [4] developed upon the open-source Java search engine, Lucene [1]. This library has several simple but useful extensions to the faceted search paradigm, which we shall use. For the purpose of this work, the two most important extensions are these:

- Instead of just counting the number of documents for each category, the library can aggregate other numeric expressions. E.g., the sum of these documents' relevance score to the query.
- The relation between a category and a document is not just binary (it is either attached to the document, or not) — it can be assigned a weight.

These capabilities are exactly what we need to produce related-people scores which are identical to the scores that the matrix approach described above would have produced: As explained above, given a query vector q, $(P^TD)q$ is a vector specifying for each person, his or her (indirect) relevance to the query. Let's rewrite this multiplication as $P^T(Dq)$. But Dq is nothing more than the vector of matching documents, specifying the relevance score of each document to the query. Looking at position i of this vector equality, we therefore discover that the indirect relevance of person i to the query (according to the matrix method) is identical to

$$\begin{aligned}\Big((P^TD)q\Big)_i &= \Big(P^T(Dq)\Big)_i \\ &= \sum_{j=1}^{\text{ndoc}} (P^T)_{ij}(Dq)_j \\ &= \sum_{j=1}^{\text{ndoc}} P_{ji} \cdot \text{score}_q(\text{document } j)\end{aligned}$$

If we remember that the relevance score is non-zero only for matching documents, and that P_{ji} is the known relation strength between document j and person i, we end up with the formula (as proposed in [2] with different justification):

$$= \sum_{\substack{\text{matching} \\ \text{documents } d}} \text{relation}(d, \text{person } i) \cdot \text{score}_q(d)$$

The extended faceted search indeed allows aggregating this sum for each person i (i.e., each category of the related people facet). relation$(d, \text{person}\, i)$ is available as the weight of the person-i category on document d, and $\text{score}_q(d)$ is available for each document because the facet aggregation starts after the document relevance scores have already been calculated.

The faceted search library of [4] contains two further extensions which are useful for our social search application:

The library allows associating with each category (in our case, person) a query-independent static score (or category *boost*). The final score of each category (person) is determined by multiplying its query dependent score with its static score. The static score of each person can be defined according to their relative popularity or authority, e.g., using the FolkRank score [7]. In our implementation, we chose to use *inverse entity frequency* (*ief*) [17]. It is defined as

$$\text{ief}(person) = \log(\frac{N}{N_{person}})$$

where N stands for the number of all documents in the system and N_{person} stands for the number of documents related to this person. Analogous to the *idf* score for terms, the *ief* score "punishes" categories that are related to many documents in general, hence are less specific to a given query.

The last faceted-search extension of interest is using the category weights to score the documents when searching for all documents in a certain category. In our social search application, this means that the top results for "All documents related to person P" will be documents which the person wrote, rather than merely commented on or bookmarked.

We've already seen that documents and people have an equal standing in our system when it comes to the search results (results of both types are returned for each query). The same is true for queries: in addition to textual queries, we can search by person (as explained in the previous paragraph), or use a combination of text and people as a query.

2.4 The Social Search Application

In the following we describe our social search implementation, based on social information gathered from IBM's intranet. From the internal Web 2.0 services in IBM, we chose the currently most used ones: *Dogear* [12], a collaborative bookmarking service used to bookmark and tag pages both within and outside the intranet; and *BlogCentral* [8], a blog service allowing all IBM employees to manage blogs within the intranet. We also used the enterprise directory application, called *BluePages*, to collect information about the IBMers who participated in Dogear or BlogCentral. At the time of writing, 15,779 employees assigned 337,345 bookmarks to 214,633 Web-pages, and wrote 67,564 blog threads.

The content we indexed for a Web page contained its title and the users' descriptions and tags as provided by *Dogear* (the actual content of the page was *not* crawled and therefore not available). For blogs, the indexed document was a blog thread, containing the blog entry, comments, and tags. For each person, we had a document containing the person's directory information (such as name, title, and group). Finally, people were connected, as facets, to the pages they bookmarked, and to their blog entries (as an author or as a commenter). Tags are connected to their related documents.

A static-score (or *boost*, in Lucene nomenclature) was given to each document based on the amount of activity around

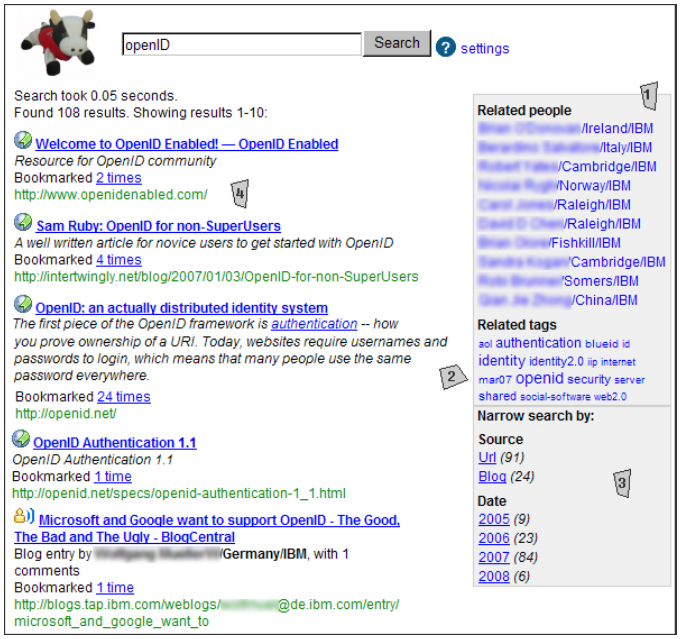

Figure 1: The social search application

it. In essence, a page which was bookmarked by many people, or a blog entry that was heavily commented or rated, is more likely to be a good search result than a document in which hardly anyone expressed interest. The actual boost used was $\log(X + 2)$, where X is the number of bookmarks, ratings and comments on that document. Our evaluation showed that this boosting significantly increased the document search precision.

The social search Web application, codenamed *Cow Search*, was made available to all users of IBM's intranet. Figure 1 shows a screenshot of the application, given the query "openID". On the left (marked by <4>) we see the most relevant documents — a mix of blogs, Web pages and personal profiles (not shown in the figure for privacy reasons). On the right <1> is the list of related people, calculated as described in the previous section. The "Related tags" tag-cloud <2> is calculated in a similar manner — each tag is a category, and the weight of the association of a tag with a document is the number of times this tag was used to describe this document. <3> shows some additional facets which aid navigation within the search results.

The list of "Related people" is not necessarily the list of IBM's experts on the topic. Experts who never bookmarked or blogged obviously cannot be retrieved by our system. Rather, the "Related people" are people that expressed interest in the topic — bookmarked a relevant document, posted a relevant blog entry, or commented on such an entry. In the next section (Evaluation) we show using a large survey that people generally agree with the system's determination of how "related" they are to various queries.

As explained in the previous section, a query can be either textual (in which case documents are matched and scored using Lucene's search algorithms), or a reference to a person (in which case the documents related to this person are returned, scored according to the relationship strength). Hybrid queries, containing both text and a reference to one or more person, are also possible.

Person and hybrid queries have a number of interesting uses. For example, giving a person as a query (or using a "everything related to this person" link) yields not only the documents related to this person, but as usual also shows related people, i.e., people who were interested in the same documents as the given person. Another useful example: For any query, if person X is a "Related person", then adding this person X to the query will find the intersection of the documents that matched the original query with the documents related to X, which is essentially the *evidence* of why person X turned up as a related person in the first place. In our application, the user can click on each of the names in the "Related people" list and choose "why this person?", to run this hybrid query. Users find this sort of evidence an important feature of the application. Another link for each person, "who is this?", displays directory information.

3. EVALUATION

Our unified search system returns both documents and people for every query. Therefore, to evaluate the quality of our system we needed to evaluate the quality of both lists.

We evaluated the document results using standard IR evaluation methodology — running 50 example queries, and having the results be judged by humans. We found the document results to be of very high quality (e.g., P@10 was 0.81). The very high precision of the top results demonstrates the capability of the social search engine to focus on good resources from the entire collection, while existing enterprise search solutions struggle with noisy datasets and have difficulties in retrieving high quality results. However, the document results are outside the scope of this workshop and therefore we will not go into details about this evaluation. Rather, in this section, we will describe in detail a large user study that we performed to evaluate the quality of the "related people" list.

From the log of queries submitted by real users to the application, we arbitrarily chose over 60 queries and ran them to receive a ranked list of 100 people for each query, using our baseline algorithm.

We then emailed all these people a list of 6-15 queries, which we defined as topics, to rate on a Likert scale of 1 to 5 whether they think the topic is relevant to them or not. We intentionally left the definition of relevance vague to address all kinds of relevance. According to replies we have received during the study, people conceived relevant to them being relevant to their work in general, their current project, their personal interests, or the interests of their team. We also did not reveal the nature of the experiment or where the topics were generated from. All of the people received along with the topic they appeared related to, a list of topics they were not found related to, thus all potentially had both relevant and irrelevant topics to rate.

We chose email rather than using a Web survey because we thought people will be more obliged to answer an email directed to them. Emails also allowed us to disassociate the application itself from the topics and hence increasing the likelihood of truthful answers not dictated by our ranking scheme. We have sent over 1400 emails for which 612 unique people replied with ratings. Those people came from 116 IBM locations in 38 countries and we assume most of them have no knowledge of each other or of our application.

From the replies we generated 8835 vote pairs of user and self-rating for 60 topics. We thus created a benchmark against which we evaluated our algorithms. To quantify our results' agreement with the benchmark, we used *normalized discounted cumulative gain (NDCG)* [9], which measures a ranked list's agreement with known relevance levels. For the NDCG calculation we used gains (0,1,3,6,10), for the 5 scale levels respectively, and the discount function used was $-log(rank + 1)$. The NDCG scores we report below are an average over the set of topics.

Despite our survey's breadth, it is to some degree biased by self-rating. Our original attempt to ask people to rate other people's interest in various topics had failed, because most respondents simply did not know enough about each other. This is why we had to ask people to rate their own interests. But self-esteem and different interpretations of the instructions inevitably lead to different people attaching different meanings to the 5 levels of the rating scale. Some people tend to over-estimate their interest in every topic, while others tend to under-estimate it. In most cases, this issue can be thought of as rating noise that is canceled out by the large number of respondents. But it can still bias some measurements. In particular, our *ief* feature is specifically designed to downplay people who over-represent their interests — contrary to those people's self-rating — and therefore we expect this survey not to measure the full value of this feature.

3.1 Results

	NDCG		
Ranking	10	20	30
count-only	0.71	0.69	0.68
sum of doc scores	0.75	0.73	0.72
+relationship weighting	0.76	0.74	0.73
+person static-score using *ief*	0.77	0.76	0.74

Table 1: The agreement of retrieved people with the system ranking of their relatedness to the searched topics, as measured by NDCG of top k results

Table 1 shows the NDCG of top k people, $k = 10, 20, 30$, measuring the agreement of retrieved people with the system's judgment of their relatedness to the searched topics. The different rows in the table show the agreements as progressively more and more features of the faceted search system were employed: The first row provides the results when people are ranked by just counting the number of their related documents (this is the "Votes" method of [11]). In the second row we ranked by summing the score of related documents to each of the people ("CombSUM" of [11]). The third row shows the results when associating different weights with the different relation types between people and documents (as in [2]); By exhaustive search we found the optimal weights to be (1, 3.1, 0) for the relation types ("tagger", "blogger", "commenter") respectively. Finally, in the last row we add the *ief* static-score for people. This row represents our full person-scoring mechanism as described in section 2.

There are several interesting insights from these results. First, the results exhibit better agreement as we consider document scores, optimal relative weights for the different relation types, and static scores for people. Second, the optimal weight for the relation between a blog entry and a commenter was found to be negative in this study (above, we used a zero weight instead, which was slightly sub-optimal).

This means that a person who comment on a blog entry is actually (slightly) *less* likely to be interested in the entry's topic than a person who did not comment on that entry. While we do not have full explanation for this result, we speculate that there are people who comment to a blog not because of its specific content but rather because of its popularity. This phenomenon should be further investigated.

4. SUMMARY

In this research, we proposed a simple yet highly effective method for a search engine to return both relevant documents and relevant people for every query. The method also allows queries to contain people instead of, or in addition to, textual terms. The list of people related to the query is calculated at search time using an enhanced faceted search engine, based on person-document relationships mined from Web 2.0 applications.

We proved that our faceted-search based method gives identical results to a more recognizable vector space model, but showed that our method has unique advantages — such as working efficiently with dynamically-changing data and taking advantage of advanced features of existing search engines (e.g., phrase search) that are not part of the classic vector space model.

We described an implementation of our method, a *social search engine* called "Cow Search" deployed in IBM's intranet. The social search engine provides several unique features not found in standard enterprise search solutions: It returns higher quality documents, as well as related people, for every query. It also allows referring to people (not just textual terms) in queries — a feature that has several interesting applications (e.g., it can be used to show the *evidence* that lead us to believe that a certain person is related to a given query).

Informal feedback from users of "Cow Search" has been very positive; Users were very pleased with the quality of both document and related-people results. We also conducted a large-scale formal evaluation. We measured the precision of the returned documents to be exceptionally high (P@10 = 0.81), and conducted a large user study, with over 600 respondents, to measure how much people agree with which topics our algorithm said were related to them. The results of this study show high agreement, and that the full method which we described is clearly better than more naive methods, such as using counting-only faceted search.

5. REFERENCES

[1] Apache Lucene. http://lucene.apache.org/java/docs/.

[2] K. Balog and M. de Rijke. Finding experts and their eetails in e-mail corpora. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 1035–1036, 2006.

[3] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, 2007.

[4] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev. Beyond basic faceted search. In *WSDM '08: Proceedings of the First ACM International Conference on Web Search and Data Mining*, 2008.

[5] K. Chakrabarti, V. Ganti, J. Han, and D. Xin. Ranking objects based on relationships. In *SIGMOD '06: Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 371–382, 2006.

[6] P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarks improve web search? In *WSDM '08: Proceedings of the First ACM International Conference on Web Search and Data Mining*, 2008.

[7] A. Hotho, R. Jäschke, C. Schmitz, and G. Stumme. Information retrieval in folksonomies: Search and ranking. In *ESWC '06: , Proceedings of the 3rd European Semantic Web Conference*, pages 411–426, 2006.

[8] 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 extended abstracts on Human factors in computing systems*, pages 2447–2452, 2007.

[9] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions of Information Systems*, 20(4):422–446, 2002.

[10] G. Luo, C. Tang, and Y. li Tian. Answering relationship queries on the web. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 561–570, 2007.

[11] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 387–396, 2006.

[12] D. R. Millen, J. Feinberg, and B. Kerr. Dogear: Social bookmarking in the enterprise. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111–120, 2006.

[13] E. Minkov, W. W. Cohen, and A. Y. Ng. Contextual search and name disambiguation in email using graphs. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, 2006.

[14] X. Wang, J.-T. Sun, and Z. Chen. SHINE: Search heterogeneous interrelated entities. In *CIKM'07: Proceedings of the 29th annual international ACM SIGIR conference on Information and Knowledge Management*, 2007.

[15] W. Xi, E. A. Fox, W. Fan, B. Zhang, Z. Chen, J. Yan, and D. Zhuang. SimFusion: measuring similarity using unified relationship matrix. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 130–137, 2005.

[16] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 107–116, 2007.

[17] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 1015–1018, 2007.

Expert Search using Internal Corporate Blogs

Pranam Kolari[†] pranam@yahoo-inc.com
Tim Finin[‡] finin@umbc.edu
Kelly Lyons[◇] kelly.lyons@utoronto.ca
Yelena Yesha[‡] yeyesha@umbc.edu

[†]Yahoo! Search Sciences, Santa Clara, CA 95054, USA
[‡]Department of CSEE, University of Maryland, Baltimore County, Baltimore, MD 21250, USA
[◇]Department of Information Studies, University of Toronto, Toronto, ON M5S 1C5, Canada

ABSTRACT

Weblogs, or blogs enable a new form of communication on the Internet. In this paper, we discuss blogs within a large corporation, and show their potential as a source of evidence to the expert search task. We describe characteristics of such blogs along multiple dimensions, and identify their utility to sub-problems within expert search. We finally discuss the use of blogs when combined with additional sources of information available within corporations.

Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Communications Applications; H.4.1 [**Information Systems Applications**]: Office Automation

General Terms

Enterprise Blogs, Expert Search

Keywords

weblogs, corporate, enterprise, expertise, blogs

1. INTRODUCTION

Many of the challenges offered by information retrieval continue to fascinate researchers. One such challenge is that of identifying and ranking the creators of information, the problem of "expert search". The immediate importance of the problem to work-force efficiency, has clearly driven focused efforts within an organizational scope [7, 8, 14, 2, 13].

Given a topic of interest, the problem consists of three interrelated sub-tasks: (i) finding relevant and authoritative sources of information, (ii) identifying and associating individuals with this information (now evidence), and (iii) combining multiple such evidence to rank individuals (now experts). Any solution could leverage diverse information sources (e.g. documents, e-mails, wikis, and distribution-lists) hosted within an organization. Though subtasks (i) and (iii) are less coupled to the nature of content, (ii) is highly tied to it. For instance, in e-mail, the association problem takes a binary form, though less direct in other sources of evidence. This simplicity, and the potential to higher precision (clear association) and recall (organizational reach), motivated early research to explore e-mail as an important source of evidence [6].

However, a dependence on e-mail has limitations, the most important being that of privacy. Recognizing this inherent limitation, combined with the generalization (to the Web at large) enabled by the use of diverse information sources, the TREC (Text REtrieval Conference[1]) expert search tasks now promote and encourage the use of pubic facing organizational content [7, 1]. This typically includes distribution-lists and many other sources of publicly available content. With this as the background, it is evident that the research community continues to explore multiple diverse sources of evidence. Each such source supplements the other, with the eventual aim of increasing overall recall (and precision) in organizational expert search.

In this paper, we present one such additional source, internal organizational (corporate or enterprise) blogs. These encompass all non-public blogs hosted within the organization on their intranets. Employees use such blogs during the course of their daily responsibilities, to share information, voice opinions, protect ownership to ideas, and to initiate discussions on issues of general interest across the organization.

Our analysis is based on internal blogs (between November 2003 and August 2006) within IBM, a global technology corporation with over 300000 employees. Blogs are published using an extended version of Roller[2], an Apache powered open source platform. Each blog is owned by an employee, or a group of employees, with a total of around 23500 blogs. These blogs host 48500 posts with a similar number of comments. Posts carry with them a timestamp, author and tags that associate content to a folksonomy[3] of topics as perceived by the author. In addition, for this study, for every employee owning a blog, information on their geographical location, and to their position and chain in the corporate hierarchy is also available.

In complementing existing sources of expert evidence, blogs provide additional benefits: (i) unlike e-mail, available for expert search from the privacy perspective, (ii) unlike other sources, providing explicit author association, timestamp and metadata, in addition to (iii) hosting topically coherent snippets of information with implicit community vote through comments. An early evidence of reduced privacy concerns is evident from its availability for researchers like us, who are external to the organization. Looking forward, we believe that content within blogs has potential similar to e-mail, and can be viewed as a social bottom-up solution to separating out shareable content from non-sharable content within an organization. We revisit our earlier work [12] on the properties of internal corporate blogs to emphasize a few of these characteristics.

The rest of this paper is organized as follows. We first show that internal blogs provide a rich source of information by discussing their growth and content properties. We next detail network prop-

[1]http://trec.nist.gov/
[2]http://rollerweblogger.org
[3]http://en.wikipedia.org/wiki/Folksonomy

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

erties and their implications. We finally discuss how certain unique characteristics of this content source could potentially enable new approaches to finding and ranking experts within an organization.

2. GROWTH AND CONTENT CHARACTERISTICS

Though less than 10% of the work-force engaged in blogs at the time of this study, their current growth suggests great long term potential. We also discuss the nature of this content, to validate how it could serve as evidence for expertise validation.

2.1 Growth of Users

At the time when it was actively tracked, the external blogosphere doubled every six months [15]. Internal blogs double at a little less than a year. Figure 1 shows the number of blogs and posts on a cumulative scale. The divergence between blogs and posts shows an interesting trend on how the blogging community is better engaging new adopters, and encouraging them to post content, hence retaining them.

To better understand how the creation of new blogs and posts trend over time we also plot the number of blogs created per month in figure 1. Two distinct spikes characterize this growth. The first, early in January 2004 was around the time when internal blogs were initiated within the organization. However, the second sharp rise around April or May 2005 was critical to the growth of blogs for two significant reasons, (i) the period following this is characterized by a dramatic increase in blog posts, and (ii) number of new blogs created every month has doubled from 500 to 1000 from before to after, suggesting that adoption was catalyzed. It turns out that at this time the organization officially embraced blogging as a communication medium and formally specified its policy and guidelines for both internal and external blogs. Evidently, having formal policies and a top-down guidance embracing blogs is key to the adoption of blogs by employees.

Driven by these organizational policy changes and high retention rate [12], we believe that the adoption is headed for continued growth, more so as the Facebook and Myspace generation enters the corporate world. Though we do not claim that blogs will support expert search task by itself, we believe its size will be significant enough to be a very important source of evidence.

2.2 Discussed Themes

We next identify themes commonly discussed within internal blogs. We use the log-likelihood approach to compare language (word usage) distributions. Informally, this measure provides a profile of content genres. We compare a random sample of content in internal blogs with that of external blogs. We first list the terms representative of internal blogs:

> *IBM, Java, code, software, team, Microsoft, Sametime, Lotus, Dogear, innovation, client, wiki, collaboration, management, social, services, customer, support, products, Websphere*

Topics of organizational nature including products, competitors and work-enviroment related issues are widely discussed in internal blogs. In contrast terms representative of external blogs are shown below:

> *journal, she, her, me, him, love, girl, lol, god, im, mom, school, shit, night, gonna, friend, tonight, eat, cry, guy, sick, happy*

Clearly, external blogs feature day-to-day activities while internal blogs focus on themes important to an organization. Many of these

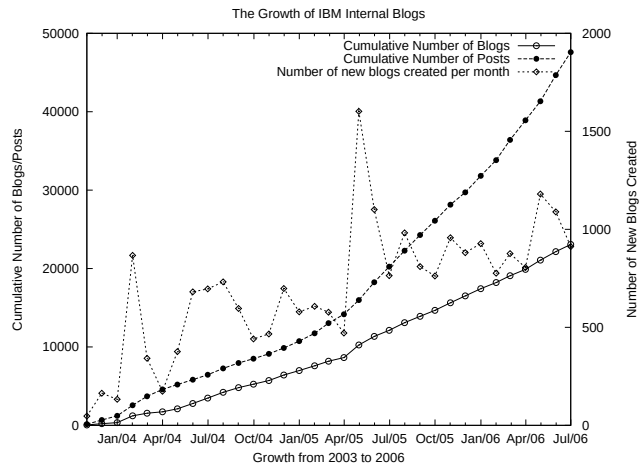

Figure 1: Growth of blogs and hosted posts has been phenomenal, with the number of blogs doubling every 10 months.

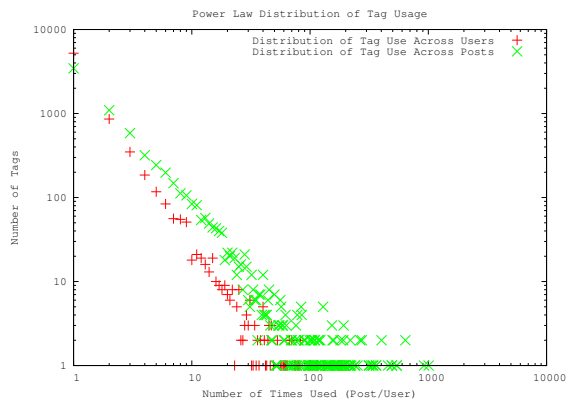

Figure 2: The distribution of tags based on their occurrence across all posts, and by the number of authors using them.

themes are topics typical of organizational expert search, suggesting that content in blogs can be a useful source of evidence.

2.3 Use of Tags

Tagging is fast becoming a common way of associating keywords (tags) to organize content. The set of all tags within a specific system or application defines a folksonomy i.e. a vocabulary of terms. We analyze to see how tags, and the concept of folksonomy is adopted by blog authors. Close to 80% of all posts are tagged, with an average of over two tags per post. However, a discussion of folksonomies is incomplete without understanding its quality [11] i.e. a folksonomy is of very little use if every user uses a distinct vocabulary of terms.

We study two attributes that have a potential bearing on quality, (i) the distribution of tags across all posts, and (ii) the distribution of tags across users making these posts. A tag provides better value to a folksonomy when used many times, and by multiple users. Figure 2 shows the distribution of tags across all posts on a logarithmic scale. The usage follows a power-law distribution indicating that a small number of tags are used with a high frequency, and a large number of them are rarely used. Similarly, the second plot in figure

2 represents the distribution of tags across users. More authors using the same tags could potentially reflect well on quality.

Since tags are less susceptible to spam in a controlled enterprise environment, the general agreement on a subset of tags suggests many common and important themes are discussed across internal blogs, again providing additional content for expert search. Though arguable that tags might add little additional information gleanable from posts [4], they can still be used as a summarization of topicality, a key attribute for expertise evidence.

2.4 Links from posts

Using posts from 2 months, we analyze how many posts feature out-links (hyperlinks), both internal and external to the organization. 60% of all posts feature out-links of one form or the other. Out of these posts, close to 70% had links to the domain of the enterprise, 50% to other domains and 22% to other internal blogs. Clearly this data point further emphasizes that *employees largely blog about themes of interest to the organization they work for*. The use of blogs, as a complementary data source, can provide useful information on authoritativeness of other sources of evidence i.e. documents discussing topics of interest to the organization that are not necessarily blogs. These characteristics could be useful in evaluating the value of new expertise evidence.

3. NETWORK CHARACTERISTICS

We next move to the study of network properties of internal blogs. Many such properties have been found useful in expert search. To materialize a social network, we generate a directed graph $G(V, E)$, where V is a non-empty finite set of vertices or nodes, and E is a finite set of edges between them. Every user u, independent of whether she owns a single blog or multiple blogs, represents a vertex in G. A directed edge e from node u to node v exists in G, if user u has commented on, or linked to, a blog post made by user v. Each such edge represents an *interaction*. We call such a graph, a *blog interaction graph*, since it reflects interactions across users through blogs. G represents a social network across all users.

We pre-process G to eliminate self-loops, to collapse multiple edges between nodes into a single edge, and to prune disconnected nodes. After pre-processing, the graph consists of 4500 nodes with 17500 edges. In the rest of this section we discuss some of the structural properties of this network, and its implications to expert search. Our analysis makes use of the JUNG[4] toolkit.

3.1 Degree Distribution

The degree distribution of a network is significant in understanding the dynamics of a network and its resilience to the deletion of nodes [3]. For every node u in G, the in-degree d_{in} and the out-degree d_{out} is computed as the number of incoming and outgoing edges respectively. The in-degree $P(d_{in})$, and out-degree distributions $P(d_{out})$ are then plotted on a log-log scale, and the power-law exponents γ_{in} and γ_{out} computed using a line fit.

The in-degree and out-degree distribution of G follows a power-law as shown in figure 3, with $\gamma_{in} = -1.6$ and $\gamma_{out} = -1.9$. This is a little lesser than their values found on the Web ($\gamma_{out} = -2.67, \gamma_{in} = -2.1$) [5], but comparable to e-mail networks ($\gamma_{out} = -2.03, \gamma_{in} = -1.49$) [9]. In the context of expert search, this scale-free property of the network shows the *resilience of the community to user attrition*. It also shows how the network of users is amenable to finding experts.

[4]http://jung.sourceforge.net/

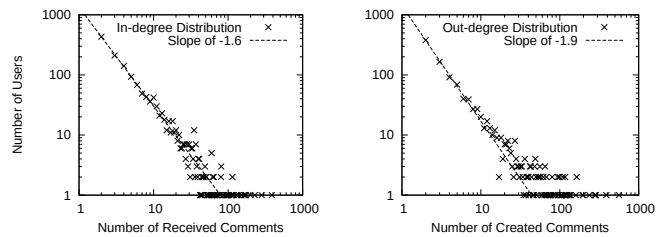

Figure 3: The in-degree of the network follows a power-law with slope -1.6 i.e. a few users generate most of the conversation. The out-degree of the network similarly follows a power-law with slope -1.9 with a few users contributing to many conversations.

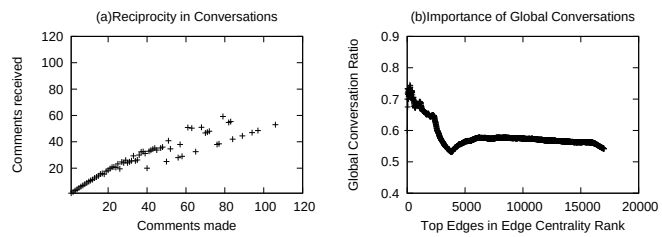

Figure 4: (a) A high correlation between in-degree and out-degree shows the reciprocal nature of blog comments. (b) A high number of cross-geography comments among high ranked central edges shows that blogs could help surface cross-geography experts.

3.2 Degree Correlation

Another interesting property of any communication medium is degree correlation. In blogs, it measures the reciprocal nature of comments i.e. *Do users who receive a number of comments, make a similar number of comments?* We plot the average out-degree of all nodes with the same in-degree. Results are shown in figure 4(a). The correlation holds for smaller degrees, but diverges randomly at higher values, possibly due to insufficient data points at such values. In general, active users in the community in addition to hosting comments on their own blog, also contribute to comments on other blogs. This has an interesting implication. It suggests that many of the popular authors (if experts) are also keen in engaging with other users within the organization. These are the experts within the organizations who could be more receptive to queries from other individuals.

3.3 Edge Betweenness Centrality

Betweenness centrality [10] measures the significance of nodes and edges as it relates to their centrality in information flow through the network. It hence forms an important measure for identifying effective word of mouth channels within a community. Many of the central nodes are key connectors within the organizations. To identify if edges that reflect interactions across geographies are central to the network, we rank edges based on their centrality, computed by finding the number of times a specific edge features in a shortest-path between every pair of nodes.

Using a ranked list of such central edges, we plot the distribution of edges that cross geographical boundaries (countries). As seen in figure 4 (b), the high ratio of such cross-geography edges

among the top ranks show the value of global interactions. Such edges form significant bridges to information dissemination across a global organization. Unlike other sources, blogs are known to be more effective in surfacing experts from such interactions.

Readers interested in many other related network properties, including those of graph ranking, are referred to our earlier work [12] on corporate blogs.

4. DISCUSSION

Motivated by many of these properties, we developed an expert search prototype for use within the organization. The application used a simple approach to topical expert search, (i) posts that serve as evidence on a topic were identified through matching tags (ii) such posts were associated to their unique authors, and (iii) these authors were ranked for expertise by comparing the aggregate number of comments to all their posts (a simplistic voting model). The tool was exposed as a tag cloud, with topical expert search limited to tags in the folksonomy. The developed application was submitted to IBM's internal hack day (which required the application be developed within a day), and was voted among top five entries by IBM employees. It was also showcased through other initiatives within IBM. Many of these resulted in feedback that can provide interesting cues to expert search, moving further. We discuss one such direction that involved tuning the comment based voting model.

In a conversation (all comments around a blog post), the relative position of employees part of the interaction, as measured through the corporate hierarchy, can be useful to understand the reach and spread of posts, and in-turn topical expertise. To support this, the employee hierarchy is modeled as a rooted named unordered tree, T. The root of the tree is the head of the organization. Each employee-manager relation is represented using a parent-child relation making managers internal nodes in the tree, and all non-managerial employees leaves.

We briefly introduce some basic tree properties. A node is an ancestor of another node u, if it exists in the shortest path from u to the root node. The height of a node u in T, denoted as $h(u, T)$ is the distance between the node u to the root of the tree, with the height of root node being zero. The Lowest Common Ancestor (LCA) of any two nodes u and v in a tree is the lowest node in T that has both u and v as descendents. We define a sub-tree $T_{LCA}^{u,v}$, as a tree rooted at the LCA of u and v and featuring only nodes and edges that are in the path from u and v to the LCA. $E(T)$ is the set of all edges in the tree T.

As opposed to only using the number of comments, the concept of *spread*, defined as the number of edges in the union of all comments around a blog post could be useful. Spread is defined as:

$$S_p(u, V) = \frac{|\bigcup_{v \in V} E(T_{LCA}^{u,v})|}{|V|}$$

Noticeably, the distribution of normalized spread across all blog posts peaks at around four [12], suggesting that conversations are high across users working in close hierarchical proximity, and less exclusive among peers, and between employees and their managers. Overall, we believe this property of conversations could signify an interesting attribute of blogs. If e-mail conversations are evidence of expertise from a 'peer' perspective, and generic documents (or mailing-lists) are evidence from a 'global' organization perspective, blogs could potentially be evidence from the 'department' as a whole.

Though the above hypothesis demands further analysis, it does point to an interesting new direction to quantify the utility of blogs. More generally, it suggests reviewing existing sources of evidence within expert search, and evaluating and accommodating new sources. While we begin to extend this work to the more generic expert search, we encourage researchers to continue exploring blogs as a useful source of evidence.

5. ACKNOWLEDGMENTS

The authors would like to acknowledge the contributions made by Stephen Perelgut and Jennifer Hawkins at IBM Toronto Software Labs, and IBM for supporting this research.

6. REFERENCES

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

[2] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, New York, NY, USA, 2006. ACM.

[3] A.-L. Barabasi. Emergence of Scaling in Random Networks. *Handbook of Graphs and Networks*, pages 69–84, 2004.

[4] B. Berendt and C. Hanser. Tags are not Metadata, but Just More Content - to Some People. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.

[5] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1-6):309–320, 2000.

[6] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531, New York, NY, USA, 2003. ACM.

[7] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *Proceedings of TREC-2005*, Gaithersburg, Maryland USA, November 2005.

[8] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. P@NOPTIC Expert: Searching for experts not just for documents. In *Poster Proceedings of AusWeb(2001)*, 2001.

[9] H. Ebel, L.-I. Mielsch, and S. Bornholdt. Scale-free topology of e-mail networks. *Physical Review E*, 66, 2002.

[10] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[11] M. Guy and E. Tonkin. Folksonomie, tidying up tags? *D-Lib Magazine*, 62(1), 2006.

[12] P. Kolari, T. Finin, Y. Yesha, Y. Yesha, K. Lyons, S. Perelgut, and J. Hawkins. On the Structure, Properties and Utility of Internal Corporate Blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2007)*, March 2007.

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

[14] D. Mattox, M. T. Maybury, and D. Morey. Enterprise expert and knowledge discovery. In *Proceedings of the HCI International '99 (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction*, pages 303–307, Mahwah, NJ, USA, 1999. Lawrence Erlbaum Associates, Inc.

[15] D. Sifry. State of the blogosphere, october, 2006.

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 dis-

cuss 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.

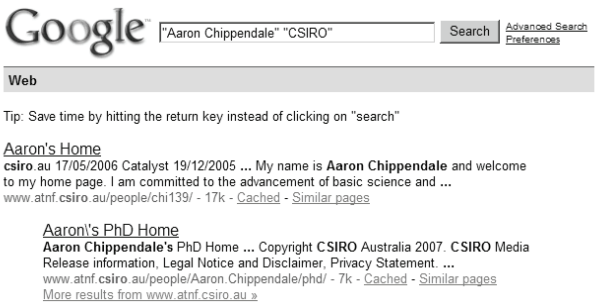

Figure 1. An example of filtered results returned by Google.

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.

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

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.

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

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 re-

sults 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 DEM-OIR 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.

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

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

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

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:

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

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),

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

	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

- **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).

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

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

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

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.

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.

[20] www.programmableweb.com/apis/
[21] tech.yahoo.com
[22] codesearch.google.com
[23] answers.yahoo.com
[24] linkedin.com/answers

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.

[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.

Multidisciplinary Expertise Retrieval with an Application in R&D Management

Choochart Haruechaiyasak Alisa Kongthon

Human Language Technology Laboratory (HLT)
National Electronics and Computer Technology Center (NECTEC)
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
choochart.haruechaiyasak, alisa.kongthon@nectec.or.th

ABSTRACT

Previous works in Expertise Retrieval (ER) mainly focused on finding people with a specific knowledge within an organization. In this paper, we propose a new challenging task called Multidisciplinary Expertise Retrieval (MULTI-ER). We define the MULTI-ER as a process of finding a group of expert candidates whose combined expertise is required to solve a multidisciplinary R&D problem. The MULTI-ER is different from the ordinary ER in two following ways. Firstly, the problem considered in the MULTI-ER is of a larger scale, and thus, requires multidisciplinary expertise from more than one person. Secondly, the scope of expert finding is not only limited within an organization, but could extend to cover people from different organizations and institutions around the world. As an illustration, a case study on the research subject of Emerging Infectious Diseases (EIDs) is used for a discussion in the context of the proposed MULTI-ER framework.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Design, Management

Keywords

Expertise retrieval, expertise identification, expert finding, expert profiling, expertise modeling, R&D management

1. INTRODUCTION

Most publicly available search engines sometimes return a long list of search results which do not exactly match the user's query. These search results are Uniform Resource Locators (URLs) which point to some specific Web pages. Therefore, such search engines only perform *document* retrieval task, rather than the actual *Information* Retrieval (IR) as many users expect. One important reason is due to the unstructured and open-domain characteristics of the Web contents. To allow the IR technique become more practical, some domain-specific IR tasks have been proposed. One of these tasks is *Expertise Retrieval* (ER) which has recently gained increasing attention among researchers in the IR community.

Previous works in the ER can be broadly classified into two groups: *expert finding* and *expert profiling* [3]. The expert finding task aims to identify a list of people who carry some certain knowledge specified by the input query. Typical approach applied for the expert finding is based on the construction of some IR models around expert candidates and topics [2, 7, 8, 9]. The expert profiling, on the other hand, focuses on identifying the area of expertise associated with a given person [4, 6]. To construct an expert profile, two types of information which can used to describe an expert are *topical* and *social* information. The topical information represents domain and degree of knowledge in which an expert possesses. The social information measures an association aspect among experts such as research project collaboration, publication co-authoring and program committee assignment.

To support the evaluation of the ER task, some related corpora have been proposed during the past few years. The first publicly available corpus is the TREC 2005 Enterprise Track [5]. The corpus consists of various contents, such as Web pages, emails, and source codes, collected by crawling on the Word Wide Web Consortium (W3C) Web site. The assigned expert search task is to identify a list of W3C people who are experts for each given topical query. The main drawback of the TREC corpus is the topical queries were directly drawn from the working groups. By using this knowledge, the models which are constructed on documents related the working groups would obviously yield good performance. This makes the TREC corpus less realistic. A more recent corpus is the CSIRO Enterprise Research Collection (CERC) which represents some real-world search activity within an enterprise [1]. The highlight of the CSIRO corpus is the use of internal staffs called *science communicators* to create some topics and perform the judgment.

The previous ER task, however, only focused on finding people with a specific expertise to solve a small-scale problem at the intra-organizational level. In this paper, we propose a new challenging task called *Multidisciplinary Expertise Retrieval* (MULTI-ER). The MULTI-ER is a process which identifies and forms a group of expert candidates whose combined expertise is required to solve a multidisciplinary R&D problem. For example, organizing a research forum to discuss on the global warming issue would require different experts with various knowledge such as scientists

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

in various fields, sociologists and policy makers.

Compared to the previous ER task, the MULTI-ER has two following differences.

- **Problem scale**: Typical ER task focuses on small and specific problems, e.g., finding programmers who are experts on Java network programming. On the other hand, the MULTI-ER considers problems which are much larger and broader, e.g., global warming, emerging infectious diseases, alternative energy resources and international terrorism. To solve these problems, multiple experts with multidisciplinary expertise are required.
- **Expert scope**: Problems considered for the ordinary ER occur within an organization, and thus, require only the internal employees. The MULTI-ER, on the other hand, considers larger-scale problems which could be interorganizational, national or even global level. Therefore, experts from many different organizations and institutions with various knowledge and expertise may be required to successfully solve the problems.

The MULTI-ER has a potential application in R&D management. Typical tasks in R&D management are, for example, organizing a forum or a meeting to discuss on a certain problem issue and forming a research workgroup to collaborate on a given project. These R&D management tasks are usually performed manually with the following steps. Firstly, the assigned problem is analyzed to identify all related topics. The next step involves searching and obtaining a list of potential candidates who are considered experts in each of the related topics. The last step is the mapping between each related topic and the candidates based on their profiles. All of the above processes require management staffs and managers who are fully trained and highly experienced.

Although the proposed MULTI-ER framework cannot fully replace humans in performing the R&D management tasks, it could provide a decision support function to improve the overall efficiency and effectiveness. Based on the MULTI-ER framework, a system could be implemented to assist and guide users in performing the tasks step by step. Many techniques in the fields such as IR, NLP and machine learning could be applied to make the process more automatic and efficient. For example, given a R&D problem as a query, the system could perform the IR and text mining tasks to automatically retrieve and extract all related keywords associated with the problem. These related keywords could be organized or clustered into a set of topics which is then verified by the users.

In next section, we present the proposed MULTI-ER framework and give a comparative discussion to the ordinary ER. In Section 3, a case study on the research subject of Emerging Infectious Diseases (EIDs) is used for a discussion in the context of the proposed MULTI-ER framework. The conclusion is given in the last section.

2. THE MULTI-ER FRAMEWORK

To support the MULTI-ER task, we propose a framework which contains different components to handle all related processes as illustrated in Figure 1. The proposed framework consists of three main components which can be explained in details as follows.

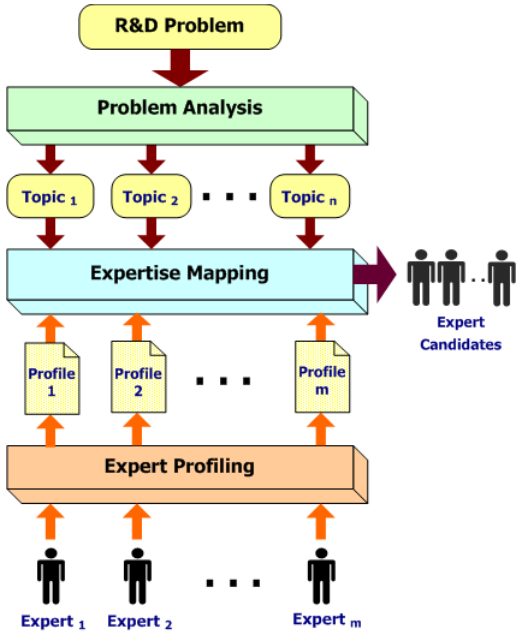

Figure 1: The Proposed MULTI-ER Framework.

2.1 Problem Analysis

The main function of the problem analysis is to analyze a given R&D problem and identify a set of n related topics. To support this process, the IR and text mining techniques could be applied. Search engines can be used to find some relevant documents on the problem. Text mining could then be applied to extract key terms related to the problem. The extracted terms could be clustered to form a set of topics. In the ordinary ER task, the problem analysis does not exist since the problem is very specific and equivalent to a small topic.

One important issue concerning the problem analysis is the information resource for supporting the process. To ensure the maximum topical coverage on a given problem, many resources should be included. Some potential resources include publication and patent databases. Most of these well-organized databases also provide some hierarchical concepts or categories which could be used to form the required set of topics. Many related IR techniques, such as query expansion and citation analysis, could also be applied to help find the relevant topics. The problem analysis is, however, difficult to evaluate since the success on solving a given problem depends on many factors besides the topical coverage. However, a group of initial experts could be asked to verify whether the set of topics meet the requirement to solve a given problem.

2.2 Expert Profiling

The expert profiling is a task which has previously been explored in the context of ER. The main goal of the expert profiling is to identify the expertise associated with an expert. The output from the expert profiling is a set of m profiles describing each of the m experts. Previous approaches in constructing expert profiles used two types of information, *topical* and *social*, to describe an expert. The topical information represents the knowledge area of an expert. The social information measures the social association

Table 1: Comparison between the ordinary ER and the proposed MULTI-ER

Factors	ER	MULTI-ER
Problem scale	Small and specific	Large and multidisciplinary
Expert scope	Within a single organization	Across multiple organizations
Data resources	Intranet and internal DBs	The Internet and outside DBs
Organizational level	Intra-organizational, e.g., W3C and CSIRO	Interorganizational, e.g., UN, WHO and FAO

between an expert to others.

Constructing expert profiles for the MULTI-ER task is different from the previous approaches proposed for the ER task. In the MULTI-ER, the problem scale is larger and the expert scope are broader than those in the ER. Therefore, the set of terms and concepts used to describe the area of expertise is extensively increased. The hierarchical concepts provided with the databases, such as the Library of Congress Classification (LCC) and the International Patent Classification (IPC), could be effectively applied to form the area of expertise.

Another important difference is the supporting information resource used for extracting the expertise. For the previous ER task, the expert scope is limited within an organization. Therefore, the information used to support the profile construction are, for example, Web pages, personal homepages, emails, blogs and Web board messages, which are available on the organization's intranet. For the proposed MULTI-ER, the information needed for building a profile must be obtained from outside of the organization. Web search engines are perhaps the main tool for gathering the information related to each expert. In addition to providing the topical information, the databases such as publications and patent records are also useful in identifying the social aspect a given expert. The social information can be analyzed and extracted from, for example, co-authoring and co-citation information.

2.3 Expertise Mapping

Once the expert profiles are available and all relevant topics to the problem are identified, the expertise mapping performs the matching between the profiles and the topics. The relationship between an expert's profile and a set of topics are one-to-many, i.e., a person could have more than one area of expertise which can be mapped into multiple topics. The output from this process is a group of candidates whose combined expertise are needed to solve the given problem.

The main design issue in the expertise mapping is the efficient ranking scheme to select an appropriate set of candidate experts to successfully solve a problem. Making the decision to form a group of candidates is not straightforward as it depends on many factors. Some important factors are experience levels of the experts and the success level of the previously performed works.

Table 1 summarizes the differences between the proposed MULTI-ER and the ordinary ER. The comparing factors are problem scale, expert scope, data resources and organization level. As mentioned in the introduction section, two main differences between the MULTI-ER and the ordinary ER are the problem scale and the expert scope. Another differences are the supporting data resources used to construct the models and the organizational level which corresponds to the problems.

Due to the larger problem scale and the broader expert scope, the MULTI-ER requires the access to external DBs and the Internet. To construct an evaluation corpus for the MULTI-ER, the potential resources could be obtained from some international organizations who deals with some multidisciplinary problem issues such as the United Nations (UN), the World Health Organization (WHO) and the Food and Agriculture Organization (FAO).

3. A CASE STUDY OF EMERGING INFECTIOUS DISEASES (EIDS)

To better understand the proposed MULTI-ER framework, we present a discussion through a case study of R&D management in Emerging Infectious Diseases (EIDs). The research subject in EIDs has currently received a lot of attentions due to the periodically reports of avian influenza outbreak. This case study aims to explore the possibility of using converging technologies which can cross discipline and contribute to the prevention and management of EIDs that are (and could become) widespread in the APEC (Asia-Pacific Economic Cooperation) region [1]. The resulting technology roadmaps will be recommended to related APEC groups, member economies, and industry for further implementing, especially to develop the technologies. The EIDs problem requires multidisplinary expertise, ranging from medical science, biotechnology, nanotechnology, material sciences, to information and communication technology. Figure 2 illustrates examples of converging technologies that can help prevent and manage EIDs.

The EIDs case study can be applied in the MULTI-ER context as follows:

- **Problem analysis**: We started analyzing the problem by gathering related information from different sources such as reports from WHO and FAO and research publications from related journals and conferences. Then we did a series of focused interview with an initial group of experts to obtain various kind of information such as the current problems and issues with EIDs and what products and services are needed in order to combat EIDs. We also applied text mining approach to help identify research topics and problems. This method could help discover topics that our initial group of experts might not be familiar with. Using various methods to analyze the problem along with the verification from the group of experts, five research topics related to combating EIDs were identified as:

  1. Bioterrorism and Surveillance System,
  2. Earth and Climate Observation,
  3. Disease Detection,

[1]EID: Roadmapping Converging Technologies for Combat Emerging Infectious Diseases, *http://www.apecforesight.org*

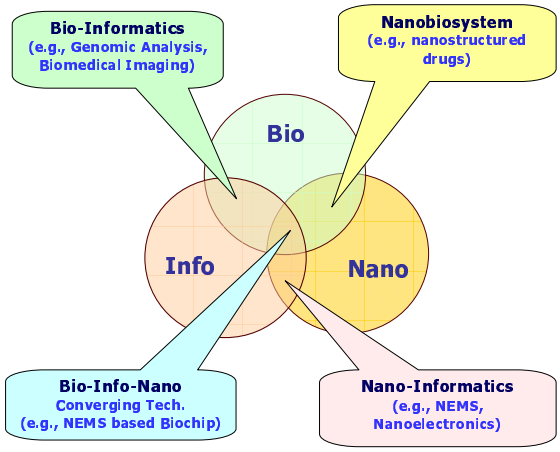

Figure 2: Multidisciplinary expertise for EIDs.

4. Disease Diagnosis,
5. Disease Identification

- **Expert profiling**: After we identified the problems and what products and services are needed in order to combat EIDs, we started identifying technologists and researchers who have expertise in such areas both locally and internationally by looking at their CVs and research papers. We also looked at their research social activities, e.g., publication co-authoring, research project collaboration, research societies they belong to.

- **Expertise mapping**: To map the identified topics with the profile of experts, we conducted three roundtable meetings among the program committee and our target experts to form a consensus.

The above example shows that the process of identifying experts for a large multidisciplinary research project depends significantly on the human experts. This could lead to some disadvantages including incomprehensive topical coverage and biased selection of expert candidates. Thus, the proposed MULTI-ER framework could provide decision support function to assist in making the overall processes more efficient and effective.

4. CONCLUSIONS AND OPEN DISCUSSION ISSUES

We proposed and gave detailed discussion on a framework for a new task called Multidisplinary Expertise Retrieval (MULTI-ER). Two main differences between the existing ER and the proposed MULTI-ER are the scope of experts and scale of problems to be solved. The MULTI-ER focuses on much larger-scale problems which could be further segmented into smaller topics. These topics are varied and thus require different experts beyond the scope of a single organization.

The MUTLI-ER introduces many new challenging research issues which can be organized into two groups: research on related techniques and development of an evaluation corpus. Some questions which must be considered on each issue are listed as follows.

- **Problem analysis**: What types of information should be considered? How to make sure that all relevant topics are included to solve the problem successfully?

- **Expert profiling**: What types of information are needed to describe the expertise of a person? How to model an expert's experience for profile construction?

- **Expertise mapping**: How to rank the expertise scores on a given topic? Should the social information be weighted more than the topical information?

- **Corpus construction**: Which organization should be considered? How to obtain the information resources to build a real-world corpus to evaluate the framework? How many topic queries should be included?

- **Performance metrics**: What types of performance measures are suitable for evaluating the MULTI-ER framework?

- **User involvement**: The experience and feedback from the users are very useful to develop a successful framework. How to introduce the framework to the people in R&D management? Which process is considered the most important for the users?

We believe that the proposed MULTI-ER is a potential and practical extension to the ordinary ER. The MULTI-ER opens up many interesting research issues which need to be discussed further.

5. REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff and A. P. de Vries. "The CSIRO enterprise search test collection," *ACM SIGIR Forum*, 41(2), pp. 42–45, 2007.

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

[3] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke and A. van den Bosch. "Broad expertise retrieval in sparse data environments," In *Proc. of SIGIR 2007*, pp. 551–558, 2007.

[4] K. Balog and M. de Rijke. "Determining expert profiles (with an application to expert finding)," In *Proc. of IJCAI 2007*, pp. 2657–2662, 2007.

[5] N. Craswell, A. de Vries, and I. Soboroff. "Overview of the TREC-2005 Enterprise Track," In *The Fourteenth Text REtrieval Conf. Proc. (TREC 2005)*, 2006.

[6] C. Macdonald and I. Ounis. "Voting for candidates: adapting data fusion techniques for an expert search task," In *Proc. of CIKM 2006*, pp. 387–396, 2006.

[7] D. Mimno and A. McCallum. "Expertise modeling for matching papers with reviewers," In *Proc. of KDD 2007*, pp. 500–509, 2007.

[8] M. Steyvers, P. Smyth, M. Rosen-Zvi and T. Griffiths. "Probabilistic author-topic models for information discovery," In *Proc. of KDD 2004*, 2004.

[9] J. Zhang, M. S. Ackerman and L. Adamic. "Expertise networks in online communities: structure and algorithms," In *Proc. of WWW 2007*, pp. 221–230, 2007.

Integrating Contextual Factors into Topic-centric Retrieval Models for Finding Similar Experts

Katja Hofmann[1], Krisztian Balog[1], Toine Bogers[2], Maarten de Rijke[1]
[1]ISLA, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
[2]ILK, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
K.Hofmann@uva.nl, K.Balog@uva.nl, A.M.Bogers@uvt.nl, mdr@science.uva.nl

ABSTRACT

Expert finding has been addressed from multiple viewpoints, including expertise seeking and expert retrieval. The focus of expertise seeking has mostly been on descriptive or predictive models, for example to identify what factors affect human decisions on locating and selecting experts. In expert retrieval the focus has been on algorithms similar to document search, which identify topical matches based on the content of documents associated with experts.

We report on a pilot study on an expert finding task in which we explore how contextual factors identified by expertise seeking models can be integrated with topic-centric retrieval algorithms and examine whether they can improve retrieval performance for this task. We focus on the task of *similar expert finding*: given a small number of example experts, find similar experts. Our main finding is that, while topical knowledge is the most important factor, human subjects also consider other factors, such as reliability, up-to-dateness, and organizational structure. We find that integrating these factors into topical retrieval models can significantly improve retrieval performance.

Categories and Subject Descriptors

H.1 [**Models and Applications**]: H.1.2 User/Machine Systems – Human information processing; 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

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Expert finding, Similar experts, Expertise seeking

1. INTRODUCTION

The goal of expertise retrieval is to support the search for experts with information retrieval technology. The need for this technology has been recognized and addressed in world-wide evaluation efforts [17]. Promising results have been achieved, mainly in the form of algorithms and test collections [2, 4]. While research in expertise retrieval has mainly focused on identifying good topical matches, behavioral studies of human expertise seeking have found that there may be important additional factors that influence how people locate and select experts. State-of-the-art retrieval algorithms model experts on the basis of the documents they are associated with, and retrieve experts on a given topic using methods based on document retrieval, such as language modeling [4]. In evaluations of these algorithms user aspects have been abstracted away. However, when a person evaluates a list of candidate experts, additional contextual factors appear to play an important role [18]—such factors include accessibility, reliability, physical proximity, and up-to-dateness.

In this paper we focus on the task of finding similar experts. We look at this problem in the context of the public relations department of a university, where communication advisors employed by the university get requests for topical experts from the media. The specific problem we are addressing is this: the top expert identified by a communication advisor in response to a given request might not always be available because of meetings, vacations, sabbaticals, or other reasons. In this case, they have to recommend similar experts and this is the setting for our expert finding task.

Our aim is to explore the integration of contextual factors into topic-centric retrieval algorithms for similar expert finding. We have two main research questions: (i) which contextual factors influence (human) similar expert finding; and (ii) how can such factors be integrated into topic-centric algorithms for finding similar experts. To answer these questions, we proceed as follows. Through a set of questionnaires completed by the university's communication advisors, we identify contextual factors that play a role in how they identify similar experts. We evaluate both topic-centric approaches and approaches with integrated contextual factors. We succeed at identifying contextual factors that play a role in this setting and show that integrating these factors with topic-centric algorithms can significantly improve retrieval performance.

The remainder of the paper is organized as follows. We discuss related work in Section 2. We discuss the organizational environment and task to which we apply our retrieval methods in Section 3. In Section 4 we describe ways of measuring topic-centric similarity of experts, which we then evaluate in Section 5. In Section 6 we analyze what additional factors play a role in human decisions on finding similar experts, which gives rise to revised similar expert finding models (in Section 7) in which we take the identified contextual factors into account. These refined models are evaluated in Section 8. We conclude in Section 9.

2. RELATED WORK

Expertise retrieval has been addressed at the enterprise track at TREC [17]. Here, retrieval is taken to the next level by focusing on retrieving entities instead of documents. Evidence from documents is used to estimate associations between experts and documents or experts and topics [4]. Two common tasks are expertise finding (given a topic, find experts on the topic) and expertise profiling (given a person, list the areas in which he or she is an expert).

A third expertise retrieval task, finding similar experts, has been formulated and addressed in [3]: an expert finding task for which a

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

small number of example experts is given, and the system's task is to return *similar experts*. Balog and de Rijke [3] define, compare, and evaluate four ways of representing experts: through their collaborations, through the documents they are associated with, and through the terms they are associated with (either as a set of discriminative terms or as a vector of term weights). Evaluation is based on the TREC 2006 enterprise search topics.

The expertise retrieval approaches discussed above focus mainly on topic-centric aspects, similar to those used for document search. However, previous research in expertise seeking has found that other factors may play a role as well. In a study of trust-related factors in expertise recommendation Heath et al. [11] find that *experience* and *impartiality* of the expert may play a role, and may additionally depend on a task's criticality and subjectivity. Borgatti and Cross [6] show that knowing about an expert's knowledge, valuing that knowledge, and being able to gain access to an expert's knowledge influenced which experts searchers would contact for help. Differences between job roles regarding the amount and motivation of expert search, as well as type of tools used indicate a possible influence of work tasks [7]. The use of social network information is expected to benefit expert search based on domain analysis [16] and users are more likely to select expertise search results that included social network information [15].

Woudstra and van den Hooff [18] focus on factors related to quality and accessibility in source selection, i.e., the task of choosing which expert candidate to contact in a specific situation. Quality-related factors include reliability and up-to-dateness of the expert, accessibility includes physical proximity and cognitive effort expected when communicating with the expert. These factors are identified in a laboratory experiment using simulated work tasks and a think-aloud protocol. The importance of individual factors is assessed through counts of how frequently they were mentioned when experts were evaluated. Quality-related factors appear to be most important while familiarity also appears to play a role.

Further initial evidence of the usefulness of individual contextual factors, such as social network information, is provided by systems that apply expertise retrieval. However, because these systems are typically not directly evaluated in terms of retrieval performance, the contribution of individual factors cannot easily be assessed. Answer Garden 2 is a distributed help system that includes an expert finding component [1]. Besides topical matches the system implements a number of heuristics found to be used in human expertise seeking, such as "staying local," i.e., first asking members of the same group or collaborators. This heuristic may be related to factors such as familiarity and accessibility. K-net is targeted at improving sharing of tacit knowledge by increasing awareness of others' knowledge [14]. The system uses information on the social network, existing skills, and needed skills of a person, which are provided explicitly by the users. Finally, SmallBlue mines an organizations' electronic communication to provide expert profiling and expertise retrieval [8]. Both textual content of messages and social network information (patterns of communication) are used. The system was evaluated in terms of its usability and utility.

3. SETTING THE SCENE

We base our study on the existing UvT Expert Collection which has been developed for expert finding and expert profiling tasks [5]. We extend the collection with topics and relevance ratings for the new task. The work task on which we focus is *finding similar experts* in the context of the public relations department of Tilburg University. The university employs six communication advisors, one responsible for the university as a whole, and one advisor for each of the faculties *Economics and Business Administration*, *Law*, *Social and Behavioral Sciences*, *Humanities*, and *Theology*. Typically, communication advisors working at a university get requests from the media for locating experts on specific topics. Such requests range from newspapers and radio shows desiring quick but informed reactions to current events, to magazine and newspaper publishers requiring more in-depth knowledge for producing special issues or articles about a certain broader theme. Locating the top expert for each request is not always trivial: the expert in question may not be available because of meetings, vacations, sabbaticals, or other reasons. In this case, the communication advisors have to recommend similar experts. This situation is the focus of our paper: what similar experts should be recommended if the top expert is not available and what factors determine what experts should be recommended?

One tool communication advisors use to find experts is *WebWijs*, a publicly accessible database of university employees who are involved in research or teaching. Each of the 1168 experts in *WebWijs* has a page with contact information and, if made available by the expert, a research description and publications list. In addition, each expert can self-assess his or her expertise areas by selecting from a list of 1491 topics, and is encouraged to suggest new topics that then need to be approved by the *WebWijs* editor. Each topic has a separate page devoted to it that shows all experts associated with that topic and, if available, a list of related topics. All of the information available through *WebWijs* was crawled to produce a test collection to evaluate algorithms for expert finding and the algorithms for finding similar experts described in this paper [5].

Another resource used for our study is the *media list*, which is compiled annually by the university's Office of Public and External Affairs and ranks researchers by media appearances, with different media types having a different influence on the score. In this scheme, media hits receive between 1 and 6 points, with mentions in local newspapers receiving 1 point and international TV appearances receiving 6 points. We considered the media rankings of the last three years (2005–2007) and collected the average and the total media score for each expert on these lists.

4. TOPIC-CENTRIC SIMILARITY

In this section we describe ways of measuring the similarity of two experts, based on two sources: (1) the (topical content of) documents authored by these experts, and (2) the expertise areas (from now on: topics) that they (optionally) selected for their expertise profile in *WebWijs*. These are baseline topic-centric retrieval approaches in that they do not take into account the contextual factors whose elicitation will be described in Section 6.

We base our approaches to measuring similarity between experts on [3], where similar approaches have been applied to similar expert finding in the W3C collection. We introduce three alternative ways of constructing the function $sim_T(e, f) \in [0, 1]$ that corresponds to the level of similarity between experts e and f. To this end, we first discuss the various expert representations and the natural ways of measuring similarity based on these representations. Finally, we consider combining the individual methods.

4.1 Representing an expert

We introduce three ways of representing an expert e. It is important to note that while these representations have been developed with an eye on the data available in our specific case (i.e., working with the data from a single specific university), they are also reasonably general, as it is not unrealistic to assume that similar sources will be available in any organization that operates at the scale of hundreds of staff members.

We use the following notation: $D(e)$ denotes the set of documents authored by expert e; $\vec{t}(d)$ is a vector of terms constructed from document d, using the TF.IDF weighting scheme; $\vec{t}(e)$ is a

term-based representation of person e, and is set to be the normalized sum of document vectors (for documents authored by e): $\vec{t}(e) = \| \sum_{d \in D(e)} \vec{t}(d) \|$. Finally, $T(e)$ is the set of topics, selected by e (from a finite set of predefined topics).

Our expert representations are then as follows.

$D(e)$ A set of documents (course descriptions and publications) associated with e.

$\vec{t}(e)$ A vector of term frequencies, extracted from documents associated with e. Terms are weighted using the TF.IDF value.

$T(e)$ A set of topics e has manually selected as his/her expertise areas.

4.2 Measuring similarity

Using the representations described above, the topic-centric similarity between experts e and f is denoted as $sim_T(e, f)$ and measured as follows. For the set-based representations ($D(e)$, $T(e)$) we compute the Jaccard coefficient. Similarity between vectors of term frequencies ($t(e)$) is estimated using the cosine distance. The three methods for measuring similarity based on the representations listed above are referred to as DOCS, TERMS, and TOPICS, respectively. Methods DOCS and TERMS are taken from [3], while TOPICS is motivated by the data made available in *WebWijs*. See Table 1 for a summary.

Table 1: Measuring topic-centric similarity.

method	data source	expert rep.	$sim_T(e, f)$
DOCS	documents	set: $D(e)$	$\frac{\lvert D(e) \cap D(f) \rvert}{\lvert D(e) \cup D(f) \rvert}$
TERMS	documents	vector: $\vec{t}(e)$	$\cos(\vec{t}(e), \vec{t}(f))$
TOPICS	expertise areas	set: $T(e)$	$\frac{\lvert T(e) \cap T(f) \rvert}{\lvert T(e) \cup T(f) \rvert}$

4.3 Combining methods

As our similarity methods are based on two sources (viz. documents and expertise areas), we expect that combinations may lead to improvements over the performance of individual methods. The issue of run combination has a long history, and many models have been proposed. We consider one particular choice, Fox and Shaw [10]'s combSUM rule, also known as *linear combination*. We combine two runs with equal weights:

$$sim_T(e, f) = 0.5 \cdot sim_1(e, f) + 0.5 \cdot sim_2(e, f), \qquad (1)$$

where sim_1 is calculated either using DOCS or TERMS and sim_2 is calculated using TOPICS. These combined runs will be referred to as DOCS+TOPICS and TERMS+TOPICS.

Similarity methods result in a normalized score in the range of $[0..1]$, but the combination could still be dominated by one of the methods. We therefore consider the linear combination in two ways:

- Score-based (S), where $sim_i(e, f)$ ($i \in \{1, 2\}$) is the raw output of the similarity method i, and
- Rank-based (R), where $sim_i(e, f) = \frac{1}{rank_i(e,f)}$ ($i \in \{1, 2\}$), and person f is returned at rank $rank_i(e, f)$ based on their similarity to expert e using method i.

5. RETRIEVAL EVALUATION

In this section we evaluate the baseline similar expert finding approaches proposed in the previous section. We start by detailing how relevance judgments were obtained (as part of a larger elicitation effort that will be described in Section 6), then we list the measures that we used for retrieval evaluation and conclude by reporting on the evaluation results.

5.1 Test set development

For our purposes, a test set consists of a set of pairs (target expert, list of similar experts). That is, in our setting, "test topics" are experts for whom similar experts need to be found.

The test topics were developed as follows. As detailed in Section 3, at Tilburg University there are six communication advisors; all participated in the experiments. For each advisor, we selected the 10 top-ranked employees from their faculty based on the media lists produced by the university's PR department; see Section 3 for details on these media lists. For one faculty the media list only contained six employees, and two employees were members of two faculties. For the university-wide communication advisor, the top 10 employees of the entire university were selected.[1] In total, then, 56 test topics were created; these included 12 duplicates, leaving us with 44 unique test topics.

For each test topic, we obtained two types of relevance judgment from the communication advisors. First, we asked the (appropriate) advisor(s) to produce one or more similar experts, together with the reasons for their recommendations and the information sources they used or would use to answer this request; the latter type of data is detailed in Section 6 below. Second, we asked the (appropriate) advisor(s) to rate the similarity of a list of 10 system-recommended experts as a substitute on a scale from 10 (most similar) to 1 (least similar). This list of 10 system-recommended experts per test topic was pooled from three different runs, corresponding to the three topic-centric baseline runs (DOCS, TERMS, TOPICS) described in Section 4. Participants were then asked to justify their rating decisions; again, see Section 6 below for details.

The expert relevance judgments were then constructed in the following way: the ratings supplied by the participants on the 10 listed experts were used as the relevance judgments for each test topic. Experts who were mentioned to be similar in part one of the questionnaire, but not in the top 10 list of part two, received the maximum relevance judgment score of 10. Experts who were not rated or not employed at the university anymore were removed. For the 12 duplicate test topics, the ratings by the two communication advisors were averaged and rounded to produce a single set of relevance judgments for each topic.

For the 12 overlapping topics, inter-annotator agreement is 75% if we only consider whether subjects selected the same top expert. Also, in half of the cases both annotators independently suggested the same expert (i.e., without seeing our suggestion list first). This relatively high agreement may indicate that subjects can easily identify a small number of similar experts. Agreement at lower ranks is difficult to establish due to low overlap between rankings (some candidates were not ranked when subjects did not feel comfortable rating a candidate), but generally appears to be much lower than at the top rank. Because of the small sample size and small number of overlapping topics we cannot draw generalizable conclusions about the reliability of our assessments.

5.2 Retrieval evaluation metrics

We used four metrics to evaluate the task of finding similar experts: ExCov, Jaccard, MRR, and NDCG. Expert coverage (ExCov) is the percentage of target experts for which an algorithm was able to generate recommendations. Because of data sparseness an expert finding algorithm may not always be able to generate a list of similar experts (for example, if the target expert did not select any expertise areas). In recommender systems evaluation, this is typically measured by coverage [12].

[1] We used the most recent version of the list that was available to us (covering 2006, while the elicitation effort took place in January 2008); this was done to ensure that the communication advisors would know the test topics and be able to suggest a similar expert.

The Jaccard similarity coefficient (Jaccard) is a statistic used for comparing the similarity and diversity of two sets. We use this measure to determine the overlap between the sets of similar experts provided by the communication advisors and by our system (irrespective of the actual rankings). Mean Reciprocal Rank (MRR) is defined as the inverse of the rank of the first retrieved relevant expert. Since communication advisors are unlikely to recommend more than one alternative expert if the top expert is unavailable, achieving high accuracy in the top rank is paramount. Given this, we will use MRR as our primary measure of performance. Normalized Discounted Cumulated Gain (NDCG) is an IR measure that credits methods for their ability to retrieve highly relevant results at top ranks. We use NDCG in our evaluation because the questionnaire participants were asked to rate the recommended experts on a scale from 1 to 10. These ratings correspond to 10 degrees of relevance, which are then used as gain values. We calculate NDCG according to Järvelin and Kekäläinen [13] using `trec_eval 8.1`.[2]

The Jaccard, MRR, and NDCG measures were computed only for experts where the similarity method resulted in a non-empty list of recommendations. In other words, "missing names" do not contribute a value of 0 to all evaluation measures. These "missing names" are instead measured by ExCov.

5.3 Results

Table 2 shows the experimental results for a total of 7 topic-centric retrieval approaches: the three similarity methods DOCS, TERMS and TOPICS listed in Table 1, plus two types of combination (DOCS+TOPICS and TERMS+TOPICS), obtained in two ways, score-based (S) and rank-based (R).

Table 2: Results, topic-centric similarity methods.

Method	ExCov	Jaccard	MRR	NDCG
DOCS	0.5227	0.1987	0.4348	0.3336
TERMS	**1.000**	0.2143	0.2177	0.3708
TOPICS	0.8409	0.3129	0.4470	0.5747
DOCS+TOPICS (S)	0.8863	0.3235	0.4529	0.5694
TERMS+TOPICS (S)	**1.000**	0.3913	0.4789*	0.6071*
DOCS+TOPICS (R)	0.8863	0.3678	**0.5422***	0.6064*
TERMS+TOPICS (R)	**1.000**	**0.4475**	0.4317	**0.6213***

A pairwise comparison of the three individual similarity methods (DOCS, TERMS, and TOPICS) indicates that these are significantly[3] different, with the exception of DOCS vs TERMS in terms of MRR. As to the combinations of runs, * marks cases where differences are significant (compared against both methods used to generate the combination).

5.4 Discussion

We see that of the three individual similarity methods, TOPICS scores best on three of the four metrics. This result is expected, because this run makes use of the human-provided self-assigned profiles. When we compare DOCS and TERMS we see that DOCS outperforms TERMS according to the MRR metric, but TERMS outperforms DOCS according to all other measures—this is in line with the findings of Balog and de Rijke [3].

Moving on to the combined methods, we see that TERMS+TOPICS is the more effective combination (according to most metrics), independent of the combination method used.

[2]The `trec_eval` program computes NDCG with the modification that the discount is always $\log_2(rank + 1)$ (so that rank 1 is not a special case).

[3]Significance is tested using a two-tailed, matched pairs Student's t-test, at significance level 0.95.

When we contrast the two combination methods (S vs. R), a mixed picture emerges. For ExCov, both score 1. For Jaccard and NDCG, the rank-based combination methods outperform the score-based one; it is the other way around for MRR.

If we look at the performance on individual topics we see that the retrieval methods used generally work well, on 23 out of the 44 test topics at least one of the methods achieves a perfect MRR score of 1.0. However, there is also a small number of topics where no relevant experts are retrieved by any of the methods. In three cases the reason is data sparseness—no topic areas or documents were available for these experts. Also, in a small number of cases, topical areas chosen by an expert are very broad (e.g., "History") so that many candidate experts are found and recommendations based on such a long candidate list are not very useful. The most interesting cases are the remaining 25% of the test topics, where documents and topic areas are available but retrieval scores are still rather low. In these cases there must be additional factors that influence human expertise recommendation decisions.

All in all, using topic-centric methods only, we manage to achieve reasonable scores, although there is clearly room for improvement. We seek to achieve this improvement by bringing in factors other than topical relevance. Before we are able to do this, however, we need to understand what these factors might be—this is our task in the following section.

6. CONTEXTUAL FACTORS IN SIMILAR EXPERT FINDING

The approaches to retrieving similar experts detailed and evaluated in the previous sections were based solely on topical relevance. In this section we seek to identify additional contextual factors that play a role in similar expert finding; in the next section we integrate some of these factors in our retrieval approach.

6.1 Methodology

Information on contextual factors was collected from (all six) communication advisors through a questionnaire; it was collected in the same study as the relevance assessments (Section 5.1). We chose this data collection method as it was deemed to require the least effort for the communication advisors whose time available for participating in the study was very limited.

The questionnaire consisted of three parts: background information, relevance assessment, and explicit rating of contextual factors. In the first part, participants were asked for information about their job function and what information sources they usually consult in their daily activities. They were also asked how often they receive requests for experts, and to give some typical examples of such requests, and how these would be addressed.

The second part of the questionnaire focused on eliciting relevance judgments for the similar experts task and factors influencing relevance decisions. We used three follow-up questions for each assessed topic in order to identify the reasons for the subjects' relevance decisions ("Why would you recommend this expert?", "Why did you rank experts in this way?", "Why did you assign the lowest score to this expert?"). Questions were formulated as open questions to allow us to discover new factors.

To compare frequencies of factor mentions to subjects' perceived importance of factors, the third part of the questionnaire asked subjects to explicitly rate the overall influence of these factors on their recommendation decisions. We used a four-point Likert-type scale and the following factors based on those identified in [18]:

Topic of knowledge the match between the knowledge of an expert and a given task

Familiarity whether and how well the subject knows the expert

Reliability the validity, credibility, or soundness of the expert's knowledge based on the expert's competence

Availability the time and effort involved in contacting the expert

Perspective the expected perspective of the expert, e.g. due to academic background

Up-to-dateness how recent the expert's knowledge is

Approachability how comfortable the subject feels about approaching the expert

Cognitive effort the cognitive effort involved in understanding and communicating with the expert and processing the obtained information

Contacts the relevance of the expert's contacts

Physical proximity how close or far away the expert is located

Saves time how much time the subject saves when contacting this expert

The questionnaire was distributed in printed form and filled out by subjects in their normal work environment and returned by mail.

6.2 Results

In this section we analyze the communication advisors' responses to part 2 of the questionnaire. We compare the identified factors mentioned in response to open questions to the explicit ratings collected in part 3, and to the findings of an earlier study [18].

The reasons subjects mentioned for relevance assessment decisions collected in part 2 of the questionnaire were transcribed and analyzed through content analysis. The responses were split into statements expressing one reason each, resulting in 254 statements. These were coded independently by two of the authors. Coding was based on the coding scheme developed in [18]; two additional factors were identified and added to the coding scheme (see below). Inter-annotator agreement was 78.3%; conflicts were resolved through discussion.

Two new factors were identified that were not present in the original coding scheme: *organizational structure* and *media experience*. Both factors can be explained by differences in tasks between the two studies. In our case the task was to recommend an expert to a media representative; in the study in [18], the experts were assumed to be sought by the subjects themselves. It appears that subjects take these task characteristics into account. Similarly, organizational structure may not have played a role in the tasks considered in [18]. In our case, this factor did play a role as candidate lists included candidates that worked in different projects, research groups, and departments within the university, held different roles (e.g., graduate student, project leader, lecturer, professor), or did not work at the university at the time the study was conducted.

Table 3 gives an overview of the frequency distribution of the resulting factors and the median rating each factor received when subjects were asked to rate these factors explicitly. *Topic of knowledge* was mentioned the most often and was mentioned by all subjects. Thus, if we assume that the frequency with which a factor is mentioned relates to the importance of the factor, then the topic is the most important. Other frequently mentioned factors are *familiarity*, and the newly identified factors *organizational structure* and *media experience*. *Physical proximity* and *saves time* were not mentioned by any subjects.

Figure 1 allows for a more detailed comparison of factors resulting from coding open responses ("implicit ratings") versus the explicit ratings subjects gave at the end of the questionnaire. There is agreement over all subjects and all measures that *topic of knowledge* is the most important factor, and *familiarity* also appears important according to both measures. Factors that appear less important according to both measures are *cognitive effort*, *saves time*,

Table 3: Example statements, frequency distribution, and explicit importance ratings (0 = *no influence*, 3 = *strong influence*) of factors mentioned. Factors marked with * were newly identified on the basis of the data.

Factor (with example statements)	Frequency (total)	Frequency (# subjects)	Median rating
Topic of knowledge (*"academic record"*, *"has little overlap with the required expertise"*, *"is only in one point similar to X's expertise"*, *"topically, they are close"*, *"works in the same area"*)	44.5%	100%	3.0
* **Organizational structure** (*"position within the faculty"*, *"project leader of PROJECT"*, *"work for the same institute"*)	24.4%	100%	n/a
Familiarity (*"know her personally"*, *"I don't know any of them"*)	17.3%	83%	3.0
* **Media experience** (*"experience with the media"*, *"one of them is not suitable for talking to the media"*)	5.5%	33%	n/a
Reliability (*"least overlap and experience"*, *"seniority in the area"*, *"is a university professor (emeritus)"*)	3.1%	33%	3.0
Availability (*"good alternative for X and Y who don't work here any more"*, *"he is an emeritus (even though he still comes in once in a while)"*)	2.4%	66%	2.5
Perspective (*"judicial instead of economic angle"*, *"different academic orientation"*)	1.2%	33%	3.0
Up-to-dateness (*"recent publications"*, *"[he] is always up-to-date"*)	0.9%	33%	3.0
Approachability (*"accessibility of the person"*)	0.4%	17%	1.5
Cognitive effort (*"language skills"*)	0.4%	17%	2.0
Contacts (*"[would] walk by the program leader for suggestions"*)	0.4%	17%	2.5
Physical proximity	0.0%	0%	0.5
Saves time	0.0%	0%	1.5

approachability, and *physical proximity*. The frequencies of *organizational structure* and *media experience* cannot be compared to explicit ratings as they were only discovered during the analysis stage.

Some factors display large disagreements in importance according to implicit and explicit rating. The largest discrepancy is found in *up-to-dateness*, which was consistently perceived as having a strong influence on expertise recommendations, but was hardly ever mentioned as a reason for a specific expertise decision. Similar differences exist between *reliability*, *availability*, and *contacts*.

We attribute the differences in importance ratings to the methodology used. A limitation of the survey format is that we do not have the possibility to clarify or encourage subjects to explore all possible factors that may have played a role in a specific decision. We therefore have to note that the frequency of factors mentioned may not give a full picture of the decisions taken and the relative importance of individual factors. For example, most candidates may be similarly reliable, and thus reliability may not be mentioned very often, even though it is very important in situations where certain candidates are more reliable than others.

The importance of these factors may also vary between faculties and between communication advisors. E.g., the Faculty of Economics and Business Administration and the Faculty of Law are both (large and) high-profile faculties that attract considerable media attention. For communication advisors of these faculties, media experience was considerably more important than for some of the smaller faculties. Faculty communication advisors also tended to recommend experts from their own faculty, whereas the university-wide advisor would recommend experts from different faculties at the same time. This suggests that the position of the communication advisor in the university's hierarchy is an important factor.

6.3 Recommendations

Based on the survey results we develop recommendations as to which contextual factors should be considered for integration in algorithms for finding similar experts in the studied task and environment. *Topic of knowledge*, *organizational structure*, *familiarity*

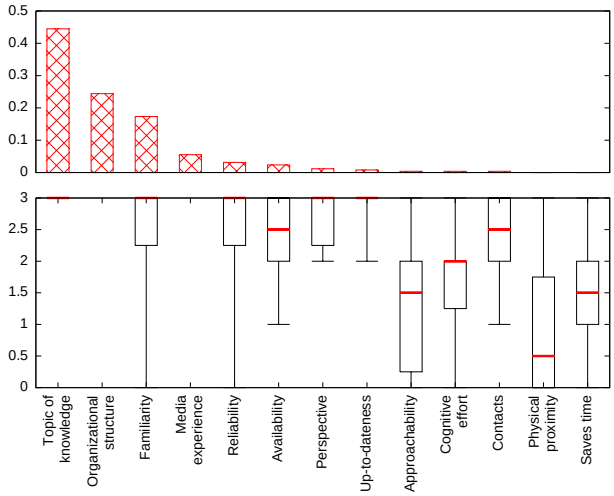

Figure 1: Frequency of implicit factor mentions (above) versus explicit ratings (below). For explicit ratings median, quartiles, minimum and maximum ratings are indicated. For *organizational structure* and *media experience* no explicit ratings are available as these factors were only identified during the analysis of the questionnaires.

and *media experience* appear promising as they received high ratings according to both implicit and explicit measures. Very interesting factors are *up-to-dateness*, *reliability*, *availability*, and *contacts*. Because of the large differences between implicit and explicit rating of these factors, results of evaluating these factors in a retrieval experiment may provide insight into the validity of the two methods used to elicit factors. *Approachability*, *cognitive effort*, *physical proximity*, and *saves time* do not appear to play a major role in the studied environment and are not discussed further.

Not all factors can be easily modeled. We discuss these aspects for each factor below; factors that will be included in the follow-up experiments in Section 7 are marked with "+" and ones that will not be considered further are marked "−".

+ **Topic of knowledge** corresponds to topic-centric similarity measures, such as the ones presented in Section 4.
+ **Organizational structure** can be implemented by taking membership in workgroups or departments into account. In our setting we have information about the organizational hierarchy down to the level of individual departments for the entire university and down to the project group level for one faculty. We can use this information to filter out experts from certain faculties or to compensate for data sparseness [5].
− **Familiarity** could be implemented in settings where social network information is available, such as patterns of email or other electronic communication (cf. related work discussed in Section 2). In our setting this type of information is currently not available.
+ Information on **media experience** can be obtained from the university's media list (cf. Section 3). These media hit counts represent a quantification of media experience and can serve for instance as expert priors.
+ **Reliability** can be modeled in various ways. For example a long publication record, or the **position** within the organization can indicate that an expert is reliable. We have access to both through the data crawled from *WebWijs*.
+ **Up-to-dateness** can be modeled by assigning higher weight to more recent documents associated with an expert, such as recent publications.
− **Perspective** is often expressed as a different angle on the same topic, such as judicial instead of economic. This suggests that looking at the organizational structure is a way of preventing too divergent perspectives. Another way of modeling this factor could be to consider co-authorship, as collaborating researchers can be expected to have a similar perspective on a topic. Currently, we do not have robust ways of estimating this factor.
− **Availability** cannot be modeled with the data currently available to us. This may be possible in systems designed to increase the effectiveness of social processes, such as awareness of co-workers' work-load [9].
+ **Contacts** similar to familiarity this factor can be modeled in systems that have access to social network information. In our case we have information about authored papers, so experts who authored many papers together are likely to be more connected. The size of their contact network can also be gleaned from these collaboration networks.

Below, we expand the topic-centric approach to similar expert finding as detailed in Sections 4 and 5 with the factors marked "+".

7. INTEGRATING CONTEXTUAL FACTORS WITH TOPIC-CENTRIC SIMILARITY

In this section we present a way of taking contextual factors into account when ranking similar experts. Ranking is based on $sim(e, f)$ and is computed as

$$sim(e, f) = p(f) \cdot sim_T(e, f), \tag{2}$$

where $sim_T(e, f)$ is the topic-centric similarity score (see Section 4), and $p(f)$ is proportional to the likelihood of expert f being recommended as similar to any other expert in question. Therefore, $p(f)$ acts as a sort of "prior probability," although here it is only requested to be a non-negative number (not necessarily a probability). In Sections 7.1–7.5 we descibe specific ways of estimating $p(f)$.

The factor *organizational structure* is not implemented as a prior but as a filtering method that limits the search space to employees from the same faculty. This approach is detailed in Section 7.6.

For the sake of simplicity, for each contextual factor addressed in this section, we demonstrate the usage of that factor in one specific way.We do not aim at being complete, nor is it our goal to push scores to the limits by carefully tuning and tailoring the methods to this specific data and test set.

7.1 Media experience

We consider the media experience of an expert according to the following formula:

$$p(f) = 1 + \log\left(1 + \sum_y media_y(f)\right), \tag{3}$$

where $media_y(f)$ is the total media appearance score of expert f for year y (see Section 3 for details about this score).

7.2 Reliability

We use the publication record of academics to estimate the degree of reliability. In principle, a long publishing record grants that

a person has valid and credible knowledge and competence. Reliability is then measured as

$$p(f) = 1 + \log(1 + \sum_y pub_y(f)), \quad (4)$$

where $pub_y(f)$ is the number of publications of expert f for year y.

7.3 Position

A second possibility for assessing an expert's reliability is their position within the university, or, more generally, the organization. E.g., a professor is more likely to be considered a reliable expert by a communication advisor than a PhD student. Here, $p(f)$ is set in correspondence to a position score associated with the staff member's title. See Table 4 for statistics over the positions of the target experts. To make use of this position information, we manually assigned $p(f)$ to each of the 19 different positions available in our data set. In this scoring $p(f)$ ranges from 0.1 to 0.9, and defaults to 0.5.

Table 4: Statistics on positions of target experts.

Position	count
Professor	29
Lecturer	7
Professor by special appointment	4
PhD student	2
Senior Lecturer	2

7.4 Up-to-dateness

Another important factor influencing the decisions of the communication advisors is the up-to-dateness of experts. An ideal candidate does not only have credible knowledge, but this knowledge is also recent. To measure this, we again use the publication records of people, but here more recent publications receive a higher weight:

$$p(f) = 1 + \log\left(1 + \sum_y w(y) \cdot pub_y(f)\right), \quad (5)$$

where $pub_y(f)$ is the number of publications of expert f for year y and $w(y)$ is the weight with which year y is taken into account. We set $w(y) = (y - 1997)/10$, where $y \geq 1997$.

7.5 Contacts

We consider only the number of co-authors, that is people that f has co-authored a publication or jointly taught a course with. Formally:

$$p(f) = 1 + \log\left(1 + coauth(f)\right), \quad (6)$$

where $coauth(f)$ is the number of distinct people with whom f has co-authored a document with or co-lectured a course.

7.6 Organizational structure

Finally, we consider the structure of the organization, which is viewed as a hierarchy of organizational units. We use only the top level of this organizational hierarchy, and consider only faculty membership information. We pursue a general scenario where a staff member may be a member of multiple faculties. The set of faculties that expert e is member of is denoted as $FAC(e)$. Unlike the other factors, organizational structure is incorporated within the retrieval process as a filtering method (not a prior). For an expert in request (e) only members of the same faculty (more precisely, experts that are members of at least one faculty that e is member of) shall be recommended as similar:

$$sim(e, f) = \begin{cases} sim_T(e, f), & FAC(e) \cap FAC(f) \neq \emptyset \\ 0, & \text{otherwise}, \end{cases} \quad (7)$$

Faculty membership information was not available for about 10% of the target experts. In those cases filtering was not applied.

Table 5: Results, combination of contextual factors and content-based similarity methods. Significant differences against the baseline are marked with *.

Method	ExCov	Jaccard	MRR	NDCG
BASELINE	1.000	0.4475	0.4317	0.6213
(1) Media experience	1.000	0.3929	0.4749	0.5967
(2) Reliability	1.000	0.3568	0.5105*	0.6113
(3) Position	1.000	0.4505	0.4317	0.6222
(4) Up-to-dateness	1.000	0.3689	0.5123*	0.6193
(5) Contacts	1.000	0.3871	0.4517	0.5956
(O) Organizational structure	0.9772	0.3607	0.4604*	0.5954*
(1) + (4)	1.000	0.3330	0.4831	0.5558*
(1) + (5)	1.000	0.3378	0.4817	0.5517*
(4) + (5)	1.000	0.3040	0.5260	0.5756*
(1) + (4) + (5)	1.000	0.2754	0.5150	0.5162*
(1) + (4) + (5) + (6)	0.9772	0.2827	0.5034	0.5277*

Table 6: Results, combination of contextual factors and content-based similarity methods. Significant differences against the baseline are marked with *.

Method	ExCov	Jaccard	MRR	NDCG
BASELINE2	0.8863	0.3678	0.5422	0.6064
(1) Media experience	0.8863	0.3725	0.4989	0.5881
(2) Reliability	0.8863	0.3508	0.5801	0.6002
(3) Position	0.8863	0.3678	0.5422	0.6064
(4) Up-to-dateness	0.8863	0.3648	0.5823	0.6119
(5) Contacts	0.8863	0.3621	0.5557	0.5930
(6) Organizational structure	0.8863	0.3363	0.5393	0.5857
(4) + (5)	0.8863	0.3281	0.5923	0.5686*

8. RESULTS AND ANALYSIS

We evaluate the contextual retrieval models introduced in the previous section using the same experimental setup as in Section 5. We apply the contextual factors on top of two topic-centric baselines. The first baseline (referred to as BASELINE) corresponds to the TERMS+TOPICS (R) run from Table 2. This run has perfect coverage (i.e., for all target experts the system is able to generate recommendations), and performs best for the Jaccard and NDCG measures. Our results using this baseline are reported in Table 5. On the other hand, one may argue that for this task MRR is the appropriate measure, since there is only one "solution," the expert to whom the media request will actually be directed; our main goal is to return this person at top rank. For this purpose we take the topic-centric run that scores best on MRR (DOCS+TOPICS (R)) as our second baseline (BASELINE2). The corresponding results are displayed in Table 6.

From Table 5 we see that with one exception (*position*) all factors improve on MRR (although the improvement is only significant for *reliability*, *up-to-dateness*, and *organizational structure*). This comes at the price of hurting the ranking at lower ranks, as is witnessed by the drops in Jaccard and NDCG. This means that these factors are indeed helpful in identifying the most similar expert. Out of these factors, the ones using the publication records of experts (*reliability* and *up-to-dateness*) seem especially helpful.

Second, when we look at Table 6, a slightly different picture emerges. *Media experience* and *organizational structure*, which helped previously, do not improve here. On the other hand, none of the differences in performance are significant. The differences are mainly due to the lack of coverage: the topics not covered by BASELINE2 are the ones that benefitted most from these factors when added to the BASELINE run. For example, for an expert without co-authorship and topic information the BASELINE still identifies some candidates based on document terms. The candidate ranked highest by the assessor was chosen due to media experience and adding this factor results in a perfect reciprocal rank score for this topic. In runs based on BASELINE2 no candidates can be found for this expert.

Finally, we experimented with combinations of individual factors; we limited ourselves to using factors that improved over the baseline and report combinations that improve over both individual runs in at least one measure. Also, out of *reliability* and *up-to-dateness*, only the latter is used, as they both rely on the publication record. Combinations improve over the baseline, but not always over the individual methods. There is considerable overlap between some factors, which indicates that more advanced methods for selecting and combining factors should be explored.

9. CONCLUSIONS

We explored the role of contextual factors in the task of finding similar experts. We started with topic-centric retrieval algorithms which were assessed in a study based on a specific work task. During relevance assessment we also collected information on contextual factors related to this task. Results of this study were used to develop recommendations for extensions of topic-centric retrieval algorithms, a number of which we implemented and evaluated. We found that the identified factors can indeed improve retrieval effectiveness.

Concerning the contextual factors that appear to play a role in human expertise finding, we find the following: while *topic of knowledge* is the most important factor, *organizational structure*, *familiarity* with the expert, and *media experience* also play a role in the setting studied. To cross-validate importance of factors we also asked subjects to explicitly rate the importance of factors on their expertise recommendation decisions. For some factors, implicit and explicit ratings corresponded well, for others, namely *up-to-dateness*, *reliability*, *availability*, and *contacts*, explicit ratings indicated high importance in contrast to implicit ratings.

As to the contextual factors for which we have appropriate data sources (and that were subsequently integrated with topic-centric retrieval models), we found that *reliability*, *up-to-dateness*, and *organizational structure* can significantly improve retrieval results as measured by MRR.

Our results indicate that identifying contextual factors and integrating them with topic-centric expertise retrieval models is indeed a promising research direction, and we hope future studies will similarly explore other expertise retrieval tasks in other environments. The method used for collecting data on contextual factors is an extension of normal relevance assessment and could be applied in other settings where the original topic creators are available for relevance assessment, such as in the TREC enterprise track.

For the retrieval models in our current work we only considered one way of implementing each factor and a limited number of ways of combining them. Some factors could not be implemented as only limited types and amounts of data were available. In the future we plan to explore other ways of integrating contextual factors with topic-centric retrieval models. The importance of contextual factors may differ between individuals, faculties, or work tasks. An interesting future direction is to address these differences through personalization. Finally, our recommendations for similar experts are solely based on the target expert, and do not take the topic of the actual request into account, as this information was not available. An appealing further direction would be to make the selection of similar experts topic dependent.

10. ACKNOWLEDGEMENTS

We are extremely grateful to the communication advisors at Tilburg University for their participation in our study: Clemens van Diek, Linda Jansen, Hanneke Sas, Pieter Siebers, Michelle te Veldhuis, and Ilja Verouden.

This research was supported by the Netherlands Organization for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, 612.066.512, and by the Dutch-Flemish research programme STEVIN under project DuOMAn (STE-09-12), and by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the IOP-MMI program of SenterNovem / The Dutch Ministry of Economic Affairs, as part of the À Propos project.

11. REFERENCES

[1] M. S. Ackerman and D. W. Mcdonald. Collaborative support for informal information in collective memory systems. *Information Systems Frontiers*, 2(3-4):333–347, 2000.

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

[3] K. Balog and M. de Rijke. Finding similar experts. In *SIGIR '07*, pages 821–822. ACM Press, 2007.

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

[5] 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. ACM Press, 2007.

[6] S. P. Borgatti and R. Cross. A relational view of information seeking and learning in social networks. *Manage. Sci.*, 49(4):432–445, 2003.

[7] K. Ehrlich and S. N. Shami. Searching for expertise. In *CHI '08*, pages 1093–1096. ACM Press, 2008.

[8] K. Ehrlich, C.-Y. Lin, and V. Griffiths-Fisher. Searching for experts in the enterprise: combining text and social network analysis. In *GROUP '07*, pages 117–126. ACM Press, 2007.

[9] T. Erickson and W. Kellogg. Social Translucence: An Approach to Designing Systems that Support Social Processes. *ACM Transactions on Computer-Human Interaction*, 7(1):59–83, 2000.

[10] E. A. Fox and J. A. Shaw. Combination of multiple searches. In D. K. Harman, editor, *Proc. TREC-2*, number 500-215 in NIST Special Publications. NIST, 1994.

[11] T. Heath, E. Motta, and M. Petre. Person to person trust factors in word of mouth recommendation. In *Proceedings of the CHI2006 Workshop Reinvent06*, 2006.

[12] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[13] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[14] S. N. Shami, C. Y. Yuan, D. Cosley, L. Xia, and G. Gay. That's what friends are for: facilitating 'who knows what' across group boundaries. In *GROUP '07*, pages 379–382. ACM Press, 2007.

[15] S. N. Shami, K. Ehrlich, and D. R. Millen. Pick me!: link selection in expertise search results. In *CHI '08*, pages 1089–1092. ACM Press, 2008.

[16] L. Terveen and D. W. Mcdonald. Social matching: A framework and research agenda. *ACM Transactions on Computer-Human Interaction*, 12(3):401–434, 2005.

[17] TREC. Enterprise track, 2005. URL `http://www.ins.cwi.nl/projects/trec-ent/wiki/`.

[18] L. Woudstra and B. van den Hooff. Inside the source selection process: Selection criteria for human information sources. *Information Processing and Management*, 2007.

Collaborative Expertise Retrieval: A Referral Approach to Finding Distributed Experts

Weimao Ke
Laboratory of Applied Informatics Research
University of North Carolina at Chapel Hill
wke@unc.edu

Javed Mostafa
Laboratory of Applied Informatics Research
University of North Carolina at Chapel Hill
jm@unc.edu

ABSTRACT

We live in a networked environment, where expertise and computing powers are highly distributed. A distributed approach to the retrieval of distributed expertise appears to be reasonable. We propose an agent simulation framework where distributed agents, representatives of information consumers, providers (experts), and referrers, learn to collaborate with each other for finding the experts. Two fundamental information organization operations, namely, clustering and classification, will be used to organize information items and to label information needs within each agent. The organized/indexed information is then mapped to the agent's perception of the society (neighbors) reinforced through machine learning. We reason why this approach is desirable and propose the investigation of: 1) whether information organization at individual levels can help expertise retrieval at the collective level; and 2) to what extent learning can facilitate the adaptive building of an efficient agent network for the finding of expertise. The proposed approach is presented as a conceptual framework. However, potentially, the implementation of the approach will provide guidance on new information and expertise retrieval models that utilize the huge distributed informational and computational resources on the Web and beyond the Web.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval—*Search process*

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Information retrieval, expertise retrieval, referral system, information filtering, information organization, agent, P2P

1. INTRODUCTION

We live in a distributed networked environment. In reality, we have different expertise, share information with each other, and ask trusted peers for information/opinions on various questions. The World Wide Web is a good example of information distribution, where Web sites serve narrow information topics and tend to form communities through hyperlink connections. Using distributed nodes to share the computational burden and to collaborate in retrieval operations appears to be reasonable.

Research on network sciences has discovered that the planet we live on is a small world with six degrees of separation [10]. That is, there are only about 5.5 social connections between any two persons in the huge population of billions. This small world phenomenon also appears in various types of networks such as the World Wide Web [1]. One implication of this is that information or expertise might be only a few degrees (connections) aways from the one who needs it. This provides the potential for distributed algorithms to traverse such a network and to find what is desired efficiently. However, the question is how we, or automatic software agents on behalf of us, can learn to find shortcuts to peers that have desired expertise.

In this work, we will propose a novel model for collaborative expertise retrieval through referrals. The task is to route (refer) information needs to experts (information providers) that have relevant information resources or expertise to satisfy the needs. We will propose the use of multi-agent simulations for the study of this distributed model and investigate: 1) whether and how information organization at individual levels can help expertise retrieval at the collective level; and 2) to what extent learning can facilitate the adaptive building of an efficient agent network for the finding of expertise. Potentially, the findings of this work will provide guidance on new information and expertise retrieval models that utilize the huge distributed computational and informational resources on the Web and beyond the Web [5].

2. RELATED WORK

Expertise Retrieval (ER) is an emerging area related to IR which recognizes the fact that individuals have distributed collections of information and expertise. In other words, it is unrealistic to assume a global collection of information at one place. Prior to the retrieval of information is the need for finding the expert(s) who potentially has the relevant information [5].

Our primary focus will be on the automatic referral for finding experts in a distributed networked environment. The paper will discuss research on Information Organization (IO) for the mapping of information needs and expertise at the individual levels. Additionally, it will utilize Machine Learning (ML) algorithms to adaptively reinforce local connections and to collectively optimize the global referral network for the efficient retrieval of expertise. Finally, it will suggest

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

the use of multi-agent technologies to simulate a distributed network for conducting experimental studies to examine the questions discussed earlier.

2.1 Distributed Info and Expertise Retrieval

Distributed IR has become a fast-growing research topic in the last few years. Recent distributed IR research has been focused on intra-system retrieval fusion, cross-system communication, decentralized P2P network, and distributed information storage and retrieval algorithms [2]. Research also concentrated on decentralized genetic algorithms for feature selection and distributed solutions for intelligent information collection and filtering.

Referral systems for expertise retrieval have attracted increasing research attention. Kautz et al. (1997) observed that much valuable information was not kept on-line for issues such as privacy. Nonetheless, this hidden information is potentially accessible through personal referrals in a social network [5]. The fact that people shared pointers to experts through word-of-mouth motivated researchers to study automated expertise retrieval systems based on referral chains [3]. The ReferralWeb was one of the early expertise retrieval systems that demonstrated promising results on referral accuracy and responsiveness [5]. In related works, software agents were used to traverse social connections for the finding of experts in an autonomously distributed manner [3, 15].

In recent years, research examined application of distributed methods to finding expertise for information filtering [12]. It was shown that, by using acquaintance lists and other collaboration strategies, learning helped distributed agents seek expertise more effectively without consuming too much communication resources. Different learning algorithms were proposed to enable effective and efficient collaboration of experts in distributed environments [6]. These studies examined various parameters for collectively finding experts on processing information. Research showed promising results on small networked communities and called for closer scrutiny on scalability.

McDonald and Ackerman (2000) acknowledged the important roles played by information mediators in organizational settings. They recognized the potential for locating expertise by making referrals and proposed an expertise retrieval system based on a range of collaborative recommendation models and behaviors [9]. The Expertise Recommender demonstrated the flexibility and usefulness of such a system for automatically locating experts in a work setting. Zhang and Ackerman (2005) studied various strategies for social network based search and found that social characteristics, in addition to graph characterisitcs, had important impacts on the searching process [16]. It also demonstrated the usefulness of simulation in distributed expertise retrieval research.

Research on distributed methods has drawn controversies over issues such as privacy and security. Some researchers reasoned that because of no centralized database, a distributed approach is more fault tolerant and less vulnerable to attacks and privacy leak [12, 3]. Others tended to disagree and argued that a distributed architecture may deploy personal information and opinions to each user, "risking exposure of information to every peer" [13, p. 317]. These questions, together with many other challenges in distributed retrieval and filtering, require continued examination.

2.2 Information Organization

Information organization is an important step toward the effective and efficient retrieval/filtering of information items. Automatic methods for information organization has been useful in centralized information retrieval operations. Although rarely discussed in distributed IR literature, it is potentially useful by building distributed indexes of expertise. We argue that, without information organization at the individual levels, it will be very difficult, if not impossible, to build orderly referral chains to expertise at the collective level.

Humans understand the world through the process of organizing concepts into an ordered group of categories. Clustering and classification, as information organization mechanisms, involve the aggregation of like-entities [8]. While clustering organizes information by grouping similar or related entities together and derives patterns (concepts) from data, text categorization, or classification, is to label texts with concepts from an existing set [14].

Text clustering and classification are fundamental functions of Information Retrieval (IR) and can be applied to various information management processes such as indexing and filtering [14]. Automatic clustering and classification, as applied in automatic information extraction and knowledge discovery, have been important research topics in Machine Learning (ML) and IR [7, 14].

The usefulness of automatic information organization for information retrieval and filtering has been extensively studied [11, 14]. Research examined the impact of information organization on automatic filtering of information, in which document classification served as an intermediate stage [11]. The proper use of classification reduced the memory size for information representation but maintained a level of granularity that could be accurately mapped to information needs.

Likewise, by making individual sets of expertise in order through information organization, peers will facilitate the distributed construction of referral chains to desired expertise. We reason that it is the ability of conceptual abstraction supported by information organization that will enable agents to understand each other's expertise. This makes possible an efficient referral network that has been demonstrated in human societies [5].

3. STUDY PROPOSAL

A multi-agent framework is useful for studying complex social and information systems. By definition, an agent is a computer program capable of autonomous action to meet its designed objectives in certain environment [4]. In multi-agent systems, agents are treated as distributed peers that have scattered intelligence and can collaborate with each other to do certain tasks. Research on information retrieval has relied upon multi-agent technologies for better understanding of collective retrieval operations in distributed environments [12, 6]. This framework also responds to the increasing computational demands for retrieval and offers a great potential for scalability.

We propose the use of multi-agent simulations for the study of expertise retrieval in a distributed networked environment. It involves referrals of information needs to experts that have matched information resources or expertise to satisfy the needs. We present the conceptual model below and elaborate on the major components of the model.

Assume that agents, representatives of information consumers, providers (experts), and referrers, live in an n dimensional space. An agent's location in the space is determined by the expertise it has. Therefore, finding experts for an information need is to route the need/query to agents in the *relevant* expertise space. To simplify the discussion, assume all experts can be characterized using two dimensions (features). Figure 1 visualizes a 2D representation of the conceptual model .

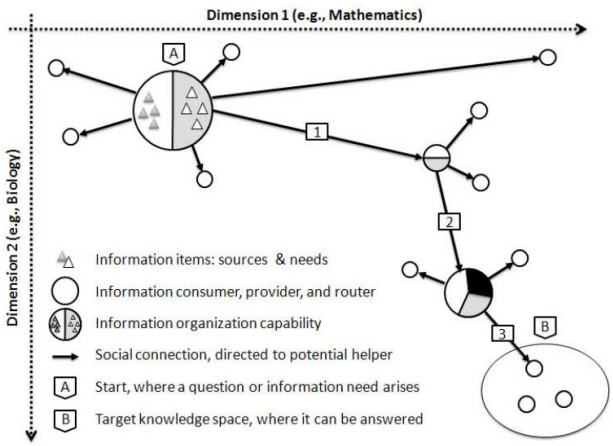

Figure 1: Agent collaboration network

As shown in Figure 1, the expertise space has two dimensions, e.g., Mathematics and Biology. Suppose Agent A has a need for expertise which is highly related to "mathematics" and "biology." That means it can be answered by agents in the space B. The problem becomes how agents in the connected society can collectively point to the right direction and find out a shortcut to the space B. We decide that, in order for us to scrutinize the dynamics of referral traffics, only one copy of each query will traverse in the network. In other words, each agent will *only* forward a query to *one* chosen neighbor. They forward it from one to another until it reaches the destination.

3.1 Organization: Index Expertise and Needs

In the expertise space, direction matters. Pointing to the right direction means the agents have to have some ability to differentiate items on certain dimensions. For instance, one should be able to tell if a query is related to mathematics or not in order to route the query properly on that dimension. This is essentially an information organization problem, which involves clustering and classification.

Firstly, an agent needs to derive patterns or concepts from information it already has through clustering. This will provide the basis of each agent's knowledge and enable the labeling of information needs. Now, when a query is posed to it, the agent will be able to tell what the query is about and assign a label to it. The label associated with the query serves as a clue for the potential referral direction.

To be realistic, each agent only has a tiny fraction of the global expertise. Hence, its information organization capability, i.e., clustering and classification, is constrained. Individual agents do not know all the dimensions of the space. In other words, each can only label the query in terms of a few dimensions–the extreme case is that one can only do *binary* classification on *one dimension*. Nonetheless, the diversity of the agent community will help if they work collaboratively. Potentially, they will refine the referral direction collectively until the query reaches the targeted expertise space.

Given that limited information within each agent, many widely appreciated classification methods, such as k Nearest Neighbor (kNN) and Support Vector Machine (SVM), require a fair amount of training data and are therefore not applicable [14]. For this study, we will use a simple centroid-based approach that has produced competitive results on a benchmark collection in a similar context [6].

3.2 Mapping Indexed Needs to Neighbors

Pointing to the right direction also requires that each agent knows which neighbor(s) should be contacted given the direction it has labeled. Therefore, there needs to be a mechanism of mapping a labeled query to a potential *good* neighbor. By *good neighbor*, we mean agents on a short path to the targeted expertise space. Sometimes, a neighbor might have the expertise to answer the query directly; sometimes, that means the neighbor can forward the query to another agent potentially closer to the desired expertise space.

Initially, of course, an agent knows nothing about its neighborhood and has to explore by trying randomly. Overtime, the agent will learn from interactions with its neighbors and get a better sense of who has (or has connections to) what expertise. In other words, a ranking function for each label can be learned from the history of interactions, which is used to predict good neighbors in the future.

In previous research [6], we explored the use of a reinforcement learning algorithm called Pursuit Learning for the automated referral to expertise on information retrieval tasks. By rewarding *successful* collaborations and penalizing *failures*, the algorithm enabled distributed agents to find experts effectively and efficiently.

Keep in mind that each agent only knows about a limited number of neighbors for a couple of reasons: first, it is rarely possible for the agents to remember every other in a huge network given their memory constraints; second, if an agent does remember all the others, it will take forever for the learning to progress–there are simply too many neighbors to explore.

3.3 Network Topology and Distributions

In the networked environment of agents, it matters how expertise is distributed and how agents connect to each other. If expertise is uniquely distributed among the agents–i.e., each agent has a unique set of information items–then the retrieval of expertise from a huge network is like finding a needle in the haystack. However, in reality, we have overlapped expertise among each other. It becomes easier to find one or some of the experts given an information need.

Another important variable in this study is the network topology, e.g., the size of the network, the in-/out-degree distributions, average path length, etc. This study assumes that every agent is connected to the network and, furthermore, the network is a small world. This is to make sure that, theoretically, all agents are only a few degrees from each other and any query can be potentially answered after a short traverse in the network.

Now the research question becomes: Given that there is indeed a shortcut (or shortcuts) to a desired expertise space,

how can agents find the shortcut(s) collectively? Additionally, as the small world phenomenon exists in a variety of networks, we do not need to arbitrarily construct such a network topology; it is something already there in reality. Seen in this light, this is a variable that we have already known and controlled.

Experimental simulations of the proposed model can be run on a scholarly communication dataset. For example, given a collection of scholarly publications, we can create a community of agents to represent scholars who authored the publications, which in turn serve as their individual expertise. A co-authorship network, presumably a small world, can be derived from the data to initialize the referral neighborhood. The simulative task for the agents could be: When assigned a paper to "read," to find the experts (ideally the authors) to help interpret the work.

To find expertise, each agent will only forward an information need (query) to one neighbor and so forth. If an agent has been contacted for a query, it will not be involved in this query again. In addition, a constraint on the maximum involvement, i.e., the maximum number of agents to be involved in each query, will be applied to all the tasks. This ensures that a query will not trouble the network for ever. We studied the maximum involvement as a variable within a small number of agents and plan to examine its impact on a large-scale agent community [6].

4. THOUGHTS ON EVALUATION

The previous sections discuss major components of the conceptual model, which involve potential independent and control variables. The dependent variables of this study are effectiveness and efficiency of expertise retrieval. We need to evaluate how accurate the found experts are or how relevant the retrieved expertise is. In addition, efficiency is also important as the model aims not to overload the network. It turns out that the efficiency evaluation will involve a couple of levels.

At the individual agent (computing node) levels, efficiency is about how fast agents can perform information organization and machine learning. It is true that these functions will consume a certain amount of computational resources. Particularly, many clustering algorithms are algorithmically complex and computationally intensive. However, this is not a primary concern in the evaluation of efficiency for the following reasons. First, each agent only runs clustering once in order to initialize its indexing space. In real situations, this can be done when a computer idles. Although classification and machine learning are needed for each query, they require less computing power than clustering does.

More importantly, the objective of the study is to take advantage of individual computing powers for indexing and learning in order to minimize network traffics for distributed expertise retrieval. Presumably, these computing resources, distributed in the network, have not been sufficiently utilized.

Summary

In this paper, we argued that a distributed architecture is desirable for the retrieval of distributed expertise in a networked environment. We proposed an automatic referral system for finding experts and elaborated on a conceptual model that can be studied using multi-agent simulations. In the model, information organization (IO) operations, namely, clustering and classification, will be used to organize information items (expertise) and to label information needs within each agent. The indexed information will then be mapped to the agent's perception of the society (neighbors) reinforced through machine learning (ML). We walked through the rationale of this model and presented initial thoughts on how to evaluate the system. Potentially, the findings of this work will provide guidance on new information and expertise retrieval models that utilize the huge distributed informational and computational resources on the Web and beyond the Web.

5. REFERENCES

[1] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[2] J. Callan, F. Crestani, and M. Sanderson. SIGIR 2003 workshop on distributed information retrieval. *SIGIR Forum*, 37(2):33–37, 2003.

[3] L. N. Foner. Yenta: a multi-agent, referral-based matchmaking system. In *AGENTS '97*, pages 301–307, New York, NY, USA, 1997. ACM.

[4] N. R. Jennings and M. J. Wooldridge, editors. *Agent technology: foundations, applications, and markets.* Springer-Verlag, Secaucus, NJ, USA, 1998.

[5] H. A. Kautz, B. Selman, and M. A. Shah. The hidden web. *AI Magazine*, 18(2):27–36, 1997.

[6] W. Ke, J. Mostafa, and Y. Fu. Collaborative classifier agents: studying the impact of learning in distributed document classification. In *JCDL '07*, pages 428–437, 2007.

[7] K. Knight. Mining online text. *Commun. ACM*, 42(11):58–61, 1999.

[8] D. D. Lewis. *Text representation for intelligent text retrieval: a classification-oriented view*, pages 179–197. Hillsdale, NJ: Lawrence Erlbaum, 1992.

[9] D. W. McDonald and M. S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In *CSCW '00*, pages 231–240, New York, NY, USA, 2000.

[10] S. Milgram. Small-world problem. *Psychology Today*, 1(1):61–67, 1967.

[11] J. Mostafa, S. Mukhopadhyay, W. Lam, and M. Palakal. A multilevel approach to intelligent information filtering: Model, system, and evaluation. *ACM Trans. Inf. Syst.*, pages 368–399, 1997.

[12] S. Mukhopadhyay, S. Peng, R. Raje, J. Mostafa, and M. Palakal. Distributed multi-agent information filtering - a comparative study. *JASIST*, 56(8):834–842, 2005.

[13] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. *Collaborative filtering recommender systems*, pages 291–324. Springer, Heidelberg, 2007.

[14] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[15] B. Yu and M. P. Singh. Searching social networks. In *AAMAS '03*, pages 65–72, 2003.

[16] J. Zhang and M. S. Ackerman. Searching for expertise in social networks: a simulation of potential strategies. In *GROUP '05*, pages 71–80, NY, USA, 2005. ACM.

Evaluating Relation Retrieval for Entities and Experts

Jianhan Zhu
University College London & The Open University
Adastral Park Campus
Ipswich, IP5 3RE, UK
j.zhu@adastral.ucl.ac.uk

Arjen P. de Vries
CWI & Delft University of Technology
Kruislaan 413, 1098SJ
Amsterdam, The Netherlands
arjen@acm.org

Gianluca Demartini, Tereza Iofciu
Leibniz Universität Hannover
L3S Research Center
Germany
{demartini, iofciu}@L3S.de

ABSTRACT
Tremendous progress has been made in terms of retrieval models and user evaluation for expert finding. From 2007, INEX provides the XML Entity Ranking track (INEX-XER) as a forum for the study of entity retrieval, a research area closely related to expert finding. Here, instead of being restricted to finding people (a particular type of entity) with a specified expertise (the topic), any type of entity related to a given topic can be the target of the retrieval system. INEX-XER 2008 proposes a novel entity relation search task, which goes beyond entity retrieval by further establishing relations between entities. Based on the connections between expert and entity retrieval, we propose to explore a tentative expert relation search task in this position paper. Our proposal shows how we can bring expert and entity retrieval research together for developing approaches that could potentially be effective for both. We expect this proposal to inspire contributions to expert finding from other research areas than information retrieval, such as semantic web, information extraction, social network analysis, virtual communities, and question answering.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: H3.1 Content analysis and Indexing; H.3.3 Information Search and Retrieval

General Terms
Experimentation, Measurement, Performance

Keywords
Relation retrieval, expert finding, entity retrieval

1. INTRODUCTION AND MOTIVATION
Many user tasks would be simplified if search engines would support typed search, and return entities instead of 'just' web pages. As an example, expert finding, i.e., retrieving people (the entity) having a specified expertise (the topic), is a key task in enterprise search and has recently attracted lots of attention from both academic and industrial communities, as evidenced by the organization of the Expert Search Task in TREC [1, 6, 13]. Since 2005, tremendous progress has been made in terms of expertise modeling, algorithms, and evaluation strategies. The goal of expert finding is to identify a list of people who are knowledgeable about a given topic. Contrary to traditional IR systems, the target of expert finding is retrieving people (named entities) instead of documents. This task is usually addressed by uncovering associations between people and topics [6].

Balog et al. [2] proposed the use of language modeling for expert finding, introducing a "Model 1" which directly represents the knowledge of an expert from associated documents and a "Model 2" which first locates documents on the topic and then finds the associated experts. Other expert finding approaches include the two-stage language model by Cao et al. [4], a generative probabilistic model by Fang and Zhai [8], a proximity-based document representation model by Petkova and Croft [10], data fusion models by Macdonald and Ounis [9], and an expert-centric language model by Serdyukov and Hiemstra [12] etc.

We observe a growing interest in extending the typed search introduced with expert finding to the retrieval of entities of other types. For example, [5] proposed the EntityRank algorithm that integrates local co-occurrence and global access information for entity search into a probabilistic estimation of entity and query association, which is quite similar to a two-stage expert finding approach. Also, INEX (INitiative for the Evaluation of XML retrieval) 2007 has started the XML Entity Ranking track (INEX-XER) to provide a forum where researchers may compare and evaluate techniques for engines that return lists of entities [7].

Expert finding and entity retrieval are closely related research areas. Since important progress has been made in expert finding since 2005, these expert finding models and techniques could also be applied to entity retrieval; as evidenced by the same random walk approach that was applied to expert finding [11] and entity retrieval [14].

In the upcoming INEX-XER 2008, we propose a new entity relation search (ERS) task investigating how well systems can not only find entities relevant to a topic but also establish correct relations between entities.

The motivation of the ERS task is that user information needs are often not satisfied with 'just' a list of entities relevant to a query, because the user would like to know more details about these entities, such as their relations with other entities, and their attributes. In a similar direction, Balog et al. [3] proposed expert profiling to complement expert finding in enterprise environment. They define the profile of an expert as her "topical profile" consisting of her skills and areas of expertise, and "social profile" in the form of her collaboration network.

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

We think that entity relation search could be applied to expert finding in terms of finding relations between experts and entities. We think that the INEX XML Entity Ranking track and TREC Expert Search task can complement each other in terms of task designs, retrieval models, and result evaluation etc. In this paper, we introduce the INEX XML Entity Ranking Track, explore its connections to expert finding, and propose an expert relation search task that can be carried out practically. In Section 2, we introduce the Wikipedia dataset for entity ranking. We give an overview of INEX-XER in Section 3. In Section 4, we propose the new Entity Relation Search (ERS) task. In Section 5, we propose expert relation search on the basis of entity relation search.

2. WIKIPEDIA DATASET FOR ENTITY RANKING

The Entity Ranking track uses the INEX Wikipedia XML collection, exploiting the category metadata about the pages to loosely define the entity sets. Given preferred categories, relevant entities are assumed to loosely correspond to those Wikipedia pages that are labeled with these preferred categories (or perhaps sub-categories of these preferred categories). Retrieval methods need to handle the situation where the category assignments to Wikipedia pages are not always consistent, and also far from complete. For example, given a preferred category 'art museums and galleries' (10855), an article about a particular museum such as the 'Van Gogh Museum' (155508) may not be labeled by 'art museums and galleries' (10855) but labeled by a sub-category of the preferred category instead, such as category 'art museums and galleries in the Netherlands' (36697). Therefore, when searching for "art museums in Amsterdam", correct answers may belong to other categories close to this category in the Wikipedia category graph, or may not have been categorized at all by the Wikipedia contributors.

3. ENTITY RANKING

Entity ranking concerns tuples of type <query, category>. The category (the entity type) specifies the type of 'objects' to be retrieved. The query (consisting of title, description, and narrative fields) attempts to capture the information need. Examples of entity ranking topics include "find European countries where I can pay with Euro", "find cities in the world where a summer Olympic game has been organized" etc. Here we can see that the set of entities to be ranked is assumed to be loosely defined by a generic category, which is often implied in the query itself, e.g., entities of type "European countries" and "cities" are desired in the above two examples, respectively. Another example of an INEX-XER topic is given in XML format.

```
<title>Impressionist art in the Netherlands</title>
<description>I want a list of art galleries and museums in the Netherlands that have impressionist art.</description>
<narrative>Each answer should be the article about a specific art gallery or museum that contain impressionist or post-impressionist art works.</narrative>
<categories>
    <category id="10855">art museums and galleries</category>
</categories>
```

We can treat expert finding as a special case of entity retrieval where we use the semantic notion of 'people' as its core category, and the query would specify 'expertise on T' for expert finding topic T. Of course, not all entity ranking queries with target category 'people' are expert finding topics; the 2007 test collection included also topics searching for presidents, tennis players and composers.

One important difference between the TREC Expert Search task and the INEX-XER bound to experts is the context and the dataset. The former focuses on the enterprise settings where the goal is to extract evidence from a dataset of e-mails and web pages, while the latter uses an encyclopedia as description of people's expertise and the queries can spread much more over all the possible topics.

4. ENTITY RELATION SEARCH

In some cases a search engine user might want to find relations between entities. In this Section we propose a new search task built on top of entity ranking. In entity relation search, we try to model a more exploratory search scenario, where people are interested in exploring the different aspects of entity ranking results. This corresponds to a view on entity relation search where the tasks are divided into an entity ranking stage, followed by the relation search stage. Given the entity ranking results, the motivation of entity relation search is here to retrieve further details about relevant entities found in entity ranking.

We call the entities found in entity ranking the main entities. Further details about the main entities are retrieved in the form of relations between each of these main entities and its related entities, which we call the target entities. The relations between main entities and target entities can be either 1 to 1, i.e., one main entity is related to one target entity, or 1 to n (n>1), i.e., one main entity is related to several target entities. These relations can be also seen as specifying (possibly multi-valued) attributes of the main entities.

Entity relation search concerns tuples of type <query, category, relation-query, target-category>. The query and category are defined in the same way as in the entity ranking task. The relation-query, given as free text, describes the desired relation between main and target entities. The relation query consists of a relation title, relation description, and relation narrative fields. The target-category specifies which category (entity type) is desired for the target entity.

The results of an entity relation search topic consist of pairs of main and target entities. For each pair of entities to be judged as a correct pair, the main entity must be judged as relevant to the original query, the main entity has to be of its correct category, the target entity is of its correct category, and the relation between them matches the relation topic.

For example, given 'Art museums and galleries' as the category and 'Impressionist art in the Netherlands' as the query topic for the main entities, 'cities' as the category for the target entities and relation query topic 'located in', we expect answer pairs like 'Van Gogh museum' and 'Amsterdam', representing the fact that the 'Van Gogh museum' is located in 'Amsterdam'.

Like in the entity ranking task, the entity types for both the main and target entities are only loosely defined by their categories 'art museums and galleries' and 'cities', respectively. Correct answers may belong to other categories close to these two categories in the Wikipedia category graph, respectively, or may not have been categorized at all by the Wikipedia contributors.

We formulate the example topic as below:

<title>Impressionist art in the Netherlands</title>
<description>I want a list of art galleries and museums in the Netherlands that have impressionist art.</description>
<narrative>Each answer should be the article about a specific art gallery or museum that contain impressionist or post-impressionist art works.</narrative>
<categories>
<category id="10855">art museums and galleries </category>
</categories>
<entity-relation>
<relation-title>located in</relation-title >
<relation-description>I want the cities where these art galleries and museums are located. </relation-description>
<relation-narrative>Each answer should be a city where a specific art gallery or museum that contain impressionist or post-impressionist art works is located. </relation-narrative>
<target-categories>
<category id="2917">cities</category>
</target-categories>
</entity-relation>

In evaluating entity relation search results, Wikipedia pages for both main and target entities are returned. The evaluator may need to read both pages in order to find evidence for judging whether their relations match the relation topic. Therefore, entity relation judgment is more complex than entity ranking judgment. After an initial pilot experiment developing some topics with assessments, we believe however that modeling relation search as an exploratory search scenario (that extends an initial ranking of the main entities to explore their attributes) alleviates this complexity sufficiently.

For evaluating the effectiveness of systems performing this task, we need to check whether they correctly identified the main entity (as in Entity Ranking), the relation, and the target entity. It is possible to extract out of the human judgments the correct (i.e., relevant) triples of the form (main entity, relation ,target entity). Similarity, it is possible to extract out of the system results the proposed (i.e., retrieved) triples. In this way we can compare relevant and retrieved results and traditional evaluation measures (such as MAP and average R-precision) can be used to measure performance of systems on entity relation search.

5. RELATION RETRIEVAL FOR EXPERTS

Our proposed entity relation search task can be applied to expert finding as well, since we may be interested in exploring further details of an expert on a search topic. We can similarly divide expert relation search into an expert search stage followed by a relation search stage. Further details about an expert are in the form of relations between the expert and other entities.

In that case, expert relation search would concern tuples of type <query, relation-query, target-category>. The query describes an expertise request, such as find experts on “semantic web”. The relation-query in form of free text describes the relation between an expert and an entity, and consists of a relation title, relation description, and relation narrative fields. The target-category specifies which category (entity type) is desired for the entity.

The results of an expert relation search topic consist of pairs of experts and entities, e.g., the relations between experts and entities of type “projects”, “organizations”, and “academic departments” can be defined as “projects-works-on”, “clients-consulted”, and “department-works-for”, respectively. Similar to entity relation search, the relations between experts and entities can be either 1 to 1, i.e., one expert is related to one entity, or 1 to many, i.e., one expert is related to several entities. For each pair of entities to be judged as a correct pair, the expert must be judged as relevant to the query, the target entity be of the correct type, and the relation between them matching the relation topic; e.g., to find “software engineering” experts and the projects they work on, for each pair consisting of person X and entity Y, we need to judge in three steps: Firstly, is person X an expert on “software engineering”? Secondly, is Y a project name? Finally, does X work on project Y?

For example, given a query topic ‘semantic web’, ‘scientific journals’ as the category for the target entities, and a relation query topic ‘published in’, the correct answers to this relation search topic will be pairs of experts and journals where each pair consists of an expert on “semantic web”, and a journal where the expert has published at least one paper.

We can formulate the example topic as below:

<title>semantic web</title>
<description>I want a list of people who are knowledge in semantic web research in my organization.</description>
<narrative>Each answer should be a person who is an expert on semantic web in my organization.</narrative>
<entity-relation>
<relation-title>published in</relation-title >
<relation-description>I want to find the journals where experts on semantic web in my organization have published papers. </relation-description>
<relation-narrative>Each answer should be a journal where an expert in my organization has published a paper.
</relation-narrative>
<target-categories>
<category id="112">scientific journals</category>
</target-categories>
</entity-relation>

In TREC2005 and 2006 Expert Search task, a crawl of the W3C website was used for expert finding [6, 13]. A predefined list of W3C related people consisting of their names and email addresses was given. Participants employed named entity recognition techniques to annotate the dataset for occurrences of these candidates. The domain for TREC2007 expert finding is the CSIRO website [1], and there was not a predefined list of candidates. Therefore, participants need to employ effective named entity recognition techniques for annotation of CSIRO related people.

In the expert relation search scenario, we envisage that either a predefined list of entities of different types would have to be provided for annotating a dataset similar to the TREC2005 and 2006 Expert Search tasks, or named entity recognition techniques should be employed to recognize entities of different categories from text like in the TREC2007 expert search task.

Of course, the cooperation of the enterprise for which to develop the collection is required. A possibly attractive alternative would be to carry out expert finding and expert relation search tasks on

the Wikipedia dataset, where a list of people and category information are readily available; an example would be to find experts on "big bang theory" in the Wikipedia dataset and find the country where each of these experts was born. Like in the entity relation search task, the category for the entities is only loosely defined, and correct answers may belong to other categories close to this category.

In evaluating expert relation search results, there are two kinds of approaches we may choose. First, ask domain experts to judge like in the TREC2007 expert search task. However, this may depend on the nature of relations and type of entities, e.g., country-of-origin of the expert does not really relate to the domain experts' domain knowledge. Second, return supporting documents for expert relations. Human evaluators judge the relations based on evidence contained in these supporting documents like in the TREC2006 expert search task [13].

Traditional evaluation measures, e.g., MAP and R-precision etc. can be used to measure performance expert relation search.

6. CONCLUSIONS

Substantial advances in terms of retrieval models and user evaluations etc. have been made in expert finding research. On the other hand, organization of the Entity Ranking track in INEX opens the door to the study of effective approaches for retrieval of entities of different types. In this paper, we explore how we can let research in expert finding and entity ranking complement each other, in particular, via our proposed relation search task for both entities and experts. We propose tentative guidelines for both entity and expert relation search tasks. We think that organization of the proposed task will help advance the research in both entity and expert retrieval by providing a platform for comparing and experimenting effective approaches for both entity and expert retrieval. A number of groups will participate in our entity relation search task in 2008. Their results on the task will provide insight into entity relation retrieval.

In the first step we design the relation retrieval task for both entities and experts as a two-stage process due to the following two reasons. Firstly, since relation retrieval task is based on entity/expert retrieval task, topic creation and user evaluation can be integrated for the two tasks. Thus topic creation and assessment can be greatly simplified. Secondly, the relationships between the two tasks can be more easily studied.

Our proposed two-stage relation retrieval task opens the door to exploring other types of relation retrieval task. One way is to focus on the relationships between main entities, e.g., finding all pairs of impressionist artists who have influenced each other or all experts in the organization who have worked together on a project etc. The challenges in this type of task can be how to formally define the relations between entities, how to evaluate the relations, and how to define the scope of such relations, e.g., how to define the "influence" relation, how to evaluate the relation, and how to know how many people "influence" each other etc.

7. ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers' comments, which have helped improve our paper.

8. REFERENCES

[1] Bailey, P., Craswell, N., de Vries, A. P., Soboroff, I. (2007) Overview of the TREC 2007 Enterprise Track (DRAFT). In Proc. of TREC 2007 Notebook.

[2] Balog, K., Azzopardi, L. and de Rijke, M. (2006) Formal models for expert finding in enterprise corpora. In Proc. of SIGIR 2006: 43-50.

[3] Balog, K., & de Rijke, M. (2007) Determining Expert Profiles (With an Application to Expert Finding). In Proc. of IJCAI 2007: 2657-2662

[4] Cao, Y., Liu, J., Bao, S. and Li, H. (2006) Research on Expert Search at Enterprise Track of TREC 2005. In Proc. of TREC 2005.

[5] Cheng, T., Yan, X., and Chang, K. C-C (2007) EntityRank: Searching Entities Directly and Holistically. In Proc. of VLDB 2007: 387-398

[6] Craswell, N., de Vries, A.P., Soboroff, I. (2006) Overview of the TREC-2005 Enterprise Track. In Proc. of TREC 2005.

[7] de Vries, A.P., Thom, J.A., Vercoustre, A-M, Craswell N., and Lalmas, M. (2007) INEX 2007 Entity Ranking Track Guidelines - V1. In Pre-Proc. of INitiative for the Evaluation of XML Retrieval.

[8] Fang, H., and Zhai, C. (2007) Probabilistic Models for Expert Finding. In Proc. of ECIR 2007: 418-430

[9] Macdonald, C. & Ounis, I. (2006) Voting for candidates: adapting data fusion techniques for an expert search task. In Proc. of CIKM 2006: 387-396

[10] Petkova, D., and Croft, W. B. (2007) Proximity-based document representation for named entity retrieval. In Proc. of CIKM: 731-740.

[11] Serdyukov, P., Rode, H., and Hiemstra, D. (2007) University of Twente at the TREC 2007 Enterprise Track: Modeling relevance propagation for the expert search task. In Proc. of TREC 2007.

[12] Serdyukov, P., and Hiemstra, D. (2008) Modeling documents as mixtures of persons for expert finding. In Proc. of ECIR 2008.

[13] Soboroff, I., de Vries, A.P. and Craswell, N. (2007) Overview of the TREC 2006 Enterprise Track. In TREC 2006.

[14] Tsikrika, T., Serdyukov, P., Rode, H., Westerveld, T., Aly, R., Hiemstra, D., and de Vries, A. (2007) Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking Using PF/Tijah. In Proc. of the 6th Initiative on the Evaluation of XML Retrieval (INEX-2007), Dagstuhl, Germany.

Author Index

NOTES