Using Citation Analysis for Finding Experts in Workgroups

Toine Bogers
ILK, Tilburg University
P.O. Box 90153, 5000 LE
Tilburg, The Netherlands
A.M.Bogers@uvt.nl

Klaas Kox
ILK, Tilburg University
P.O. Box 90153, 5000 LE
Tilburg, The Netherlands
K.Kox@uvt.nl

Antal van den Bosch
ILK, Tilburg University
P.O. Box 90153, 5000 LE
Tilburg, The Netherlands
Antal.vdnBosch@uvt.nl

ABSTRACT

We compare expert finding approaches that use and combine different types of expertise evidence: content-based expert finding using academic papers, and expert finding using a social citation network between the documents and authors. We evaluate our approaches on a test collection that represents the research output of a typical average-sized academic workgroup. We find that expert finding using static rankings achieves the same performance as a query-dependent approach. Of the different approaches, the most effective method of performing expert finding in an academic workgroup is ranking workgroup members by citation indegree.

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

Algorithms, Measurement, Performance, Experimentation

Keywords

Expertise search, expert finding, citation analysis, data fusion

1. INTRODUCTION

There is an increasing belief that enterprise search is a vital tool for meeting the demands of the global marketplace. *Expert search* is considered a crucial component of an effective enterprise search system. A successful expert search system helps an organization address two important tasks, as signaled by Maybury [26]: *expert finding* and *expert profiling*. Expert finding is the task of locating individuals or communities knowledgeable about a specific topic. A complete and up-to-date overview of the experts related to a topic, task, or assignment can for instance aid an organization in rapidly recruiting an operational team to respond to a new market opportunity or threat. Expert finding involves analyzing communications, publications, and activities. It should also include the ability to rank them on multiple dimensions such as qualifications, availability, experience, and reputation.

The term expert profiling, first coined by Balog et al. [2], encompasses all activities related to assessing expertise, such as classifying and quantifying individual expertise and the expertise of entire organizational units, and validating the breadth and depth of that expertise. A successful expert profiling system would also allow organizations to identify changes in expert profiles of individuals and organizational units [26].

In general, three different sources of information for expertise attribution can be identified within organizations [26]:

- Content-based evidence is one of the most prevalently used sources of expertise in expert finding research, typically including documents and e-mails authored by employees. Homepages, resumes, and shared folders in a file system can also be used as content-based evidence of expertise.
- Organizations are made up of a variety of social networks. We assume that people who interact are likely to share expertise. Evidence of these interactions can be found in the organization structure, but also in e-mail flow, usage of software libraries, and bibliographic information. Records of information exchange in these networks provide evidence of expertise.
- A third type of evidence for expertise is activity-based: how much time did an employee spend on a project, and what are the search and publication histories of employees.

In this paper we focus on the problem of expert finding. In particular, we investigate the impact of combining two different sources of expertise—content-based and social networks—on expert finding in an average-sized academic workgroup. The research output of such a workgroup provides a stream of content-based evidence in the form of papers and technical reports. In addition, we can also benefit from the rigorous citation culture in academia. We assume that highly cited papers are indicative of the expertise of its authors on the topics covered by those papers. We investigate the combination of this evidence with content-based methods. To our knowledge, citation analysis and content-based expert finding techniques have not yet been compared and combined; this is the contribution of the current paper.

In the following section we describe the relevant related work in expert finding and citation analysis, and the intersection of the two fields. Section 3 describes our methodology, experimental setup, and evaluation. In Sections 4 and 5 we describe our experiments with different types of expertise evidence separately. We combine these approaches in Sections 6 and 7. We conclude our paper with a discussion of the outcomes of the present study, and formulate goals for future work in Sections 8 and 9.

DIR-2008 April 14-15, 2008, Maastricht, the Netherlands.

2. RELATED WORK

2.1 Expert finding

Early large-scale approaches to expert finding came in the form of constructing and querying databases containing representations of the knowledge and skills of an organization's workforce. These systems tended to delegate the responsibility and workload to the employees, giving them the task to create and maintain adequate descriptions of their own continuously changing skills [26].

This disadvantage prompted a shift to expert finding techniques more supportive of the natural expertise location process [27], and more automatic approaches to expert finding such as the one by Campbell et al. (2003). They performed expert finding on e-mail collections of two different organizations, comparing a content-based approach with a graph-based augmented approach and reporting that the latter outperformed the purely content-based approach [11].

Arguably, the key development boost for the field of expert finding and expert profiling has been the introduction of the Enterprise track in TREC 2005. From its inception the track included an Expert Finding task, that triggered rapid advances in the field of expertise retrieval, in terms of modeling, algorithms, and evaluation methods.

Participants in the 2005 and 2006 TREC Enterprise tracks validated their work using the W3C test collection, a 2004 crawl of the World Wide Web Consortium website [13]. This collection—330,037 documents, adding up to 5.7GB, with a list of 1,092 candidate experts—contains not only web pages but also numerous mailing lists, technical documents, and other kinds of data that represent the day-to-day operation of the W3C. For the Enterprise track of TREC 2007, a new test collection was used: the CSIRO collection with 370,715 documents, totaling 4.2 GB, with a list of 3,678 candidate experts [1]. Other collections representing different types of organizations have been created, such as the UvT Expert Collection [4]. The work described in this paper is performed on a small subset of this collection (see Section 3.1 for more details).

Expert finding—identifying a list of people who are knowledgeable about a given topic—is usually approached by uncovering salient associations between people and topics [13]. The co-occurrence of a person with topics in the same context is commonly assumed to be evidence of expertise of that person on those topics. The majority of expert finding approaches can be divided into either *document-centric* or *candidate-centric* approaches. In the candidate-centric approach to expert finding, each expert is represented by a profile that is constructed from the expertise evidence associated with that expert. A simple way of doing this would be concatenating all documents associated with an expert into a single profile document for that expert. In a document-centric approach, the first step is retrieving documents or other forms of expertise evidence relevant for the query and then associating those retrieved documents with the different experts. Many different retrieval models have been used for both expert finding methods, as well as many extensions to existing retrieval models originally developed for common document retrieval, such as (pseudo-)relevance feedback, query expansion, using passage-level evidence, and re-ranking using static rankings [13, 33].

Approaches that combine different forms of evidence—such as using static rankings for re-ranking purposes—are especially interesting with regard to the topic of this paper. To our knowledge, the first to do so were the aforementioned Campbell et al. when they found that a graph-based approach performed better than a pure content-based approach [11]. Chen et al. (2006) took a similar approach while investigating social networks found in the mailing lists in the W3C corpus [12]. They used PageRank [28] to rank experts on centrality, and a revised version of the HITS algorithm [23] for submitting their runs. They compared this with a two-stage model that combined relevance with co-occurrence, and found that HITS performed significantly worse. They explain that the root cause for the lack of success is the specific nature of mailing list networks, which allow for reciprocal links to be added to the network much easier that the typical web link network, or citation network. Kolla et al. (2006) used a similar HITS-based reranking approach and reported marginal but insignificant improvements [24]. Bao et al. (2006) achieved similar results by using PageRank [5]. In contrast, the approach taken by Zhu et al. (2006) to use Google rankings turned out to be an ineffective way of improving performance [37]. Serdyukov et al. (2007) modeled the search for experts as a multi-step propagation of relevance through a hyperlinked network of relevant documents and found improvements over a one-step model [32]. Outside of TREC, another effort to use network analysis for expert location was made by Zhang et al. (2007), who used a set of network-based ranking algorithms, including PageRank and HITS, to identify expert users of a Web-based programming community [36]. They found these algorithms did not outperform simpler algorithms for expert finding.

In sum, earlier work on static rankings does not seem to yield a satisfactory answer to the question whether using static ranking techniques such as HITS and PageRank helps or hurts expert finding performance. A possible reason might be that the networks that were analyzed—mailing lists and intranet pages—are not related (enough) to expertise; they may lack uniform and overt signs of the expertise of the individuals posting emails or adding web pages.

A related research topic that shares many similarities with expert finding is automatically routing submitted papers to reviewers in conferences [6, 14, 15, 35]. All of these approaches use the sets of papers written by the individual reviewers as content-based expertise evidence for those reviewers to match them to submitted papers. The most extensive work was done Yarowksy et al., who performed their experiments on the papers submitted to the ACL'99 conference [35]. They compared both content-based and citation-based evidence for allocating reviewers and found that combining both types resulted in the best performance.

2.2 Citation analysis

Citation analysis involves assessing the research performance of individual scholars, scholarly journals, and research groups, departments, and institutions. Analyzing bibliographic networks has a rich history: the first citation indexes were developed by Eugene Garfield in the 1950s. Garfield (1979) also pioneered the use of these indexes in assessing the popularity and impact of specific articles, authors, and publications [17].

As mentioned in the previous section, we assume the degree to which a paper (or a set of papers about a topic) is cited, to be a good indicator of expertise. We are therefore interested in bibliometric indicators that help to identify the important elements in a citation network, more specifically, well-cited papers and authors. The classic example of such a bibliometric indicator is the so-called *impact factor*. Pioneered by Garfield's Institute for Scientific Information in the 1960s, the impact factor was meant to be an objective measure of the reputability of a journal [17].It is defined as the average number of citations—or average *indegree*—per article a journal receives over a two-year period.

The original impact factor formulation does not distinguish between citations: citations from journals with a high impact have the same weight as citations from low impact journals. Pinski et al. (1976) were the first to suggest a recursive impact factor to remedy this [29], with several others proposing related approaches, such as Bollen et al. (2006) who proposed using the PageRank algorithm [10, 28]. Examples of journal rankings using the PageRank algorithm can be found, for instance, on the Eigenfactor.org website[1].

In our expert finding situation we focus on a single workgroup. We therefore only cover a subset of the citation network of the workgroup's research field. We do not have the impact factors for every journal and conference proceedings. Lacking these for now, we use the indegree count for each document and author to calculate the importance of authors and documents in the network. In addition, we wish to use PageRank as a way of calculating a recursive impact factor.

Garfield, among others, has warned against using impact factors to measure the productivity of individual scientists, arguing that different scholarly disciplines can have very different publication and citation practices and that there is "wide variation from article to article within a single journal" [18]. However, we believe that the homogeneous research focus of our evaluated workgroup alleviates this problem to some extent. Furthermore, using a recursive algorithm for calculating the impact factor—such as PageRank—can also help alleviate this. We therefore decided to use these two bibliometric indicators to determine the importance of documents and authors: standard citation indegree, and PageRank scores.

Another measure that has been proposed as a way of estimating an individual researcher's impact is the so-called *Hirsch number* (or *h*-index) [20]. A scholar has a Hirsch number of *h* if he has published *h* papers that have been cited *h* times or more. We could not test this measure as an expertise estimator because it is better at distinguishing between scientists within an entire field than within a workgroup; we do not have access to the full network of the research fields.

There is some related work at the intersection of information retrieval and citation analysis. One of the first investigations into the usefulness of citations for document retrieval was performed by Salton (1963), who found a significant correlation between text-based document similarity and citation overlap similarity between documents [30]. Another obvious related example is the PageRank algorithm for ranking web pages, developed by Page et al. (1998), inspired by ideas from citation analysis. It has been successfully used to improve Web retrieval performance, for instance, by producing document priors or re-ranking retrieval results [28]. Some specific search engines for scholarly literature have been developed, most notably Google Scholar [21] and CiteSeer [19]. In general, such specific search engines perform 'normal' document retrieval, re-ranking the results by indegree (citation) count.

Drawing from the principle of polyrepresentation, Larsen combined text-based retrieval techniques with citation analysis, but found no significant improvements over a bag-of-words baseline [25]. More recently, Strohman et al. (2007) tested many seemingly useful measures descriptive of the citation network, but found that only combining text-based retrieval with the graph-based Katz measure significantly improved performance [34]. Finally, Fujii (2007) also combined text-based patent retrieval with the PageRank probabilities of citations between patents and found small but significant improvements in recall [16]. Overall, there seems to be a tendency for citation analysis to yield small improvements over normal text-based retrieval approaches.

3. METHODOLOGY

3.1 Data & experimental techniques

The test collection we use in our experiments is the ILK collection[2], previously used in expertise-based re-ranking of document retrieval [7, 8]. It contains 147 publications (titles, abstracts, and full text) of current and ex-members of the ILK Research Group. The paper topics focus mainly on machine learning applied to issues in language engineering and linguistics. The collection contains 80 natural language queries with exhaustive relevance judgments. Each query has 3.6 relevant documents on average, contained 12 terms on average, and 6 terms on average after stopword filtering . There are 89 unique authors in the collection with an average of 2.7 authors per document. An example of one of the queries is "*How do you use machine learning for named entity recognition?*". We used the small ILK collection because it is a realistic description of an academic workgroup in terms of research output and because all 147 documents contained one or more references to other papers.

The ILK collection, originally created for document retrieval experiments, is expanded to an expert finding collection based on the assumption that authors of a document possess expertise on the document topic. If a query has, say, three relevant documents, then the relevant experts for that query are the authors of those three documents. This resulted in an average of 5.1 experts per query. This assumption is similar to the one made in constructing the W3C collection. There, members of a workgroup were automatically considered experts on the topic of the workgroup, irrespective of experience or other qualifications. Both assumptions carry the inherent risk of overgeneralizing the assignment of expertise, as authorship does not always imply expertise in the topic of the document. To check for this potential overgeneralization factor, we gathered explicit expert rankings for 10 of the 80 queries by asking members of the ILK workgroup [9]. This is not sufficient for statistically reliable conclusions, but it should give us some indication of how our expert finding methods perform.

The ILK test collection is in fact a subset of the UvT Expert Collection, which does have explicit expert relevance judgments. However, we could not make use of the richer information, because only 10% of the ILK authors have a profile in the UvT Expert Collection (e.g. because they are not affiliated with Tilburg University anymore). See Section 8 for a discussion of the ramifications of our choice for this collection.

We collected the citation information using Google Scholar[3]. For each ILK document we collected the title and author information of all documents that *cited* that ILK document and all documents *cited by* that ILK document. We did not go beyond 1 level of cited and citing documents because of time limitations. We crawled citation information about 3,205 citing and cited-by documents. We did not filter out self-citations. Using a third party search engine such as Google Scholar means that we might miss citations that the engine itself missed because of misparsed citations, but no citation parsing method is perfect. On average, each ILK document was cited 15.7 times and cited 21.0 documents itself.

3.2 Evaluation

All of the expert finding work in TREC has been evaluated using binary relevance judgments. We performed both binary evaluation using the standard information retrieval measures of Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR).

3.3 Experimental design

We tested the effectiveness of the two types of evidence—the content of the academic papers and the citation network they are a part of—separately before combining them. Sections 4 and 5

[1] http://www.eigenfactor.org/

[2] Freely available at http://ilk.uvt.nl/toine/∼ilk-collection/

[3] http://scholar.google.com

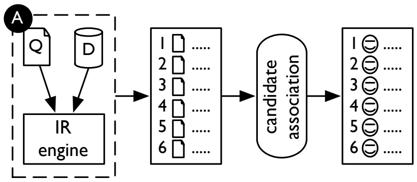

Figure 1: Two-stage content-based expert finding

describe our experiments with the two separate types of evidence, while Section 6 describes the way we combined runs. In Section 7 we tested another way of using citation analysis for expert finding: using co-citation information to re-rank the different expert finding runs. Finding 'similar' authors using co-citation analysis could help boost similar experts that were glossed over for some reason in the original expert finding runs. We report on the results of the binary evaluation after each round of experiments in Sections 4-7 separately.

4. CONTENT-BASED EXPERT FINDING

For our content-based expert finding approach we used a document-centric approach: first, we find documents relevant to a query, and then we associate the experts to the retrieved documents to produce a ranking. This two-stage approach to expert finding has been found to outperform candidate-centric approaches [3]. Instead of modeling these two stages directly as Balog et al. did, we used an existing retrieval toolkit for the first stage and used its output as input for the second stage. Figure 1 illustrates this content-based approach.

We implemented the document retrieval stage using the Lemur Toolkit[4] and tried out different combinations of retrieval algorithms: language modeling using Kullback-Leibler divergence with Jelinek-Mercer smoothing (**JM**) and Dirichlet smoothing (**DIR**); probabilistic retrieval using the Okapi BM25 term weighting scheme (**OKAPI**), and the Vector Space model using TF·IDF term weighting (**TFIDF**). We tested these different models on two different versions of the ILK collection: one with the titles and the full document text (**FULL**) and one with only the titles and abstracts (**ABSTRACT**). The different text fields were not weighed differently. Table 1 shows the results of these preliminary experiments.

Table 1: First stage document retrieval experiments. Best scores are printed in bold.

	ABSTRACT		FULL	
Model	**MAP**	**MRR**	**MAP**	**MRR**
JM	**0.7211**	**0.8825**	0.6830	0.8173
DIR	0.7096	0.8684	0.6766	0.8450
TFIDF	0.6754	0.8164	0.5784	0.7027
OKAPI	0.7035	0.8688	0.2492	0.3053

The language modeling approaches significantly outperformed the other models except Okapi on the **ABSTRACT** collection[5]. Performance on the **ABSTRACT** collection was always significantly better than performance on the **FULL** collection. There was no significant difference between the **JM** and **DIR** models. We selected the best-performing language modeling approach using Kullback-Leibler divergence with Jelinek-Mercer smoothing as our first stage model for document retrieval.

4.1 Document-author association

After retrieving the relevant documents for each ILK query, the scores assigned to each relevant documents needed to be associated with the authors of those documents to produce. Associating authors is not always simple: if there is no clear authorship to a document, then other methods are needed for estimating the association strength from, for instance, the occurrence of names in the documents [3]. In our situation this probability can be unambiguously estimated because authorship of research papers is perfectly documented. However, we did need to distinguish between how much each retrieved document contributed to a person's expertise.

In preliminary experiments we tested different methods of attributing document relevance scores to authors. With our unambiguous document-author associations, simply summing the scores would be equal to the method used by Balog et al. (2006). We also experimented with other methods, such as averaging the set of scores for each author. Space restrictions keep us from reporting those results here, but the best performing association method corrected for the returned document's ranks by summing relevance scores of each retrieved document ($score(q_k, d_i)$) by the ranks at which they were retrieved ($rank(q_k, d_i)$). The best results were achieved by moderating the rank influence by taking the 2 logarithm of the rank (see below). Not moderating resulted in significantly worse performance in all cases. Equation 1 represents our method of candidate association. In this equation, the $assoc(a_t, d_i)$ component of the equation is equal to 1 when author a_t authored document d_i and 0 if not. For each author separately we then summed the moderated m document relevance scores for the entire query q_k to produce the expertise score $expertise(a_t, q_k)$.

$$expertise(a_t, q_k) = \sum_{i=1}^{m} \left(assoc(a_t, d_i) \cdot \frac{score(q_k, d_i)}{\log_2(rank(q_k, d_i))} \right) \quad (1)$$

4.2 Expert finding results

Table 2 shows the results of our two-stage content-based approach to expert finding (**run A**). It achieved a MAP score of 0.4311 and a MRR of 0.7859 when evaluated on the set of 80 queries. If we evaluate on the set of 10 queries with real expert relevance judgments, MAP is higher at 0.5435 and MRR is perfect at 1.000. This content-based run serves as the baseline against which we compare the citation-based approach and the combined approaches of the next three sections.

Table 2: Results of the content-based expert finding run on both query sets

	80 queries		**10 queries**	
	MAP	**MRR**	**MAP**	**MRR**
run A	0.4311	0.7859	0.5435	1.0000

5. CITATION-BASED EXPERT FINDING

As mentioned in Section 2.2, we investigate two different indicators of the importance of elements in a social network: indegree and PageRank. We assume that the importance of a node is correlated with the expertise represented by that node. In this section, we

[4] http://www.lemurproject.org

[5] All comparisons done at $p < 0.05$ unless noted otherwise.

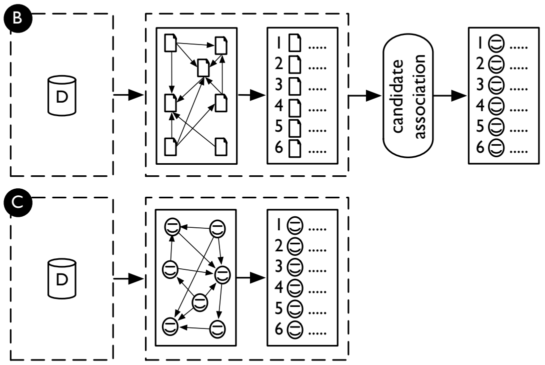

Figure 2: Citation-based expert finding

describe our two query-independent approaches to using the social network to locate experts. Focusing solely on the social network as a source of expertise results in a static ranking that is the same for each query, and is actually more of a representation of the general distribution of authority in the workgroup.

Citation networks can be defined in two different ways: as a document network and as an author network. Table 3 contains some statistics about the two networks when induced from the ILK test collection.

Table 3: Indegree and outdegree statistics of two interpretations of the ILK citation network: as a document network and as an author network.

	document network	author network
avg. indegree	15.7	503.0
avg, outdegree	21.0	489.3
max. indegree	190	1274
max. outdegree	178	1133

We calculated the static rankings of both networks using both normalized indegree (**IN**) and PageRank (**PR**). Figure 2 illustrates these two approaches. In the first run, we used the document citation network to generate a static author ranking and associated candidates to the documents as described in the previous section to generate a ranked list of experts (**run B**). Second, we directly generated this ranked list from the author citation network (**run C**). Table 4 shows the results of these two runs when ranking is performed on indegree and PageRank.

Evaluating on the 80 queries (where expertise is automatically linked to authorship), PageRank seems to produce slightly better rankings than indegree on the author citation network (run C) in terms of MAP, but this is not statistically significant. For the document citation network (run B) the situation is reversed, and also not significant. On the subset of 10 queries (with manually assigned expertise judgments) the relationships between indegree and PageRank are in the opposite directions on MAP and not significant either. All but one of the citation-based approaches do outperform the content-based approach on MRR slightly. The document citation network run using PageRank is clearly the worst performing of all citation-based runs. Besides this run, none of the MAP or MRR scores in Table 4 were significantly better or worse than the baseline approach, although all static rankings achieved scores surprisingly close to the baseline. The difference between using PageRank or indegree to produce the static ranking is also not significant, aggregated over these runs. In all but one case the scores on the set of 10 queries are higher than on the set of 80 queries.

Table 4: Results of the pure citation-based expert finding approaches. Best scores are printed in bold.

	80 queries		10 queries	
	MAP	MRR	MAP	MRR
run B – IN	0.4262	0.7925	0.4413	**1.0000**
run B – PR	0.3923	0.7478	**0.4761**	**1.0000**
run C – IN	0.4294	**0.7952**	0.4541	**1.0000**
run C – PR	**0.4341**	0.7941	0.4312	**1.0000**
baseline	0.4311	0.7859	0.5435	1.0000

When using PageRank to calculate the importance of the elements in the network, there is an additional parameter *d* that influences the calculation of the PageRank scores. The original PageRank algorithm models the behavior of a random surfer following link on the Web. Since surfers will not keep following links forever, *d* is the probability that the surfer will jump to random other page [28]. In our data we do not have a network of Web pages and links between them. To still make use of the *d* parameter, the analogy could be drawn between following hyperlinks and following the citation from a starting document to other documents (or authors) to find experts. Well-connected and well-cited documents and authors have a greater probability of being reached, which is analogous to the PageRank situation. Yet, it is hard to imagine a situation where randomly jumping to another expert would help in finding a relevant expert for that specific topic. This suggests that the teleport component of PageRank might not serve any purpose in expert finding. Our results point in this direction as well. We tested different values of *d* in our PageRank calculations: 0.15, 0.5, 0.85 (the 'default' value), and 1.0. A *d* of 1.0 effectively means there is no random jumping to other documents. We found that increasing *d* also increased MAP and MRR. We therefore set *d* to 1.0 in all our experiments with PageRank.

6. COMBINING EVIDENCE

There are many different ways in which different retrieval runs can be combined, as investigated by, among others, [22]. We restricted our combination experiments to only taking the product of the *expertise* scores from two combined runs. Using this, we tested 8 different combinations of runs A, B, and C. Figure 3 illustrates the different combinations. Run D involves combining the static document ranking with the content-based document retrieval run, following by the standard candidate association. Run E has content-based document retrieval as its starting point and after the candidate association the resulting list of experts is combined with the static author ranking run to produce a new run. Run F uses the static rankings in both stages. Each of the runs' static rankings can be calculated using indegree or PageRank, bring the total number of combinations up to 2 + 2 + (2 × 2) = 8.

Table 5 shows the results of these 8 runs. In general, there is no convincing evidence that combining the two types of expertise evidence in a workgroup setting improves performance. Combining content-based expert finding with only the static author ranking performs better than the baseline in both MAP and MRR, evaluated on the set of 80 queries, and all of the runs perform slightly better in MRR, but none of these improvements are significant. Any decreases in performance were also not significant. Finally, once again there is no significant difference between using indegree or PageRank to produce the static rankings, aggregated over all con-

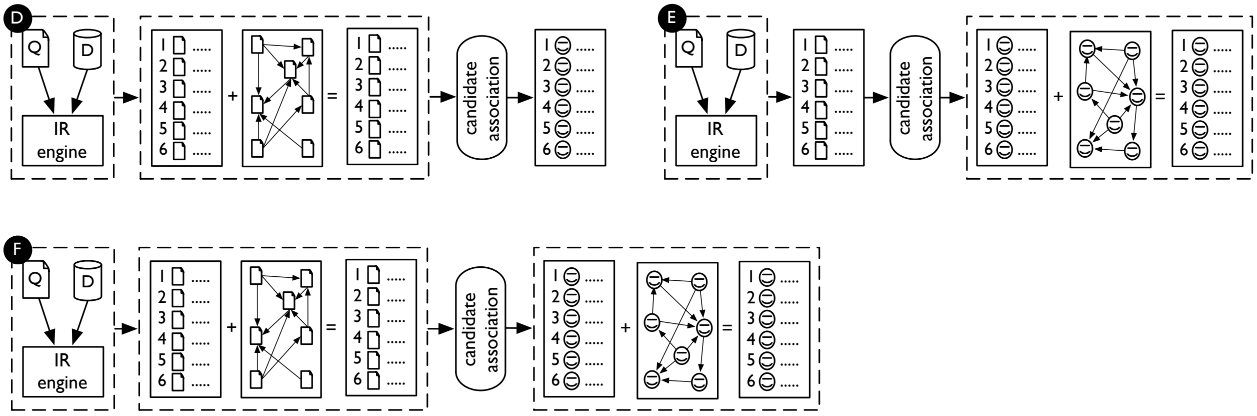

Figure 3: Combined expert finding approaches

cerned runs. In all but one case the scores on the set of 10 queries are higher than on the set of 80 queries.

Table 5: Results of the combined expert finding runs. Best scores are printed in bold.

	80 queries		10 queries	
	MAP	MRR	MAP	MRR
run D – IN	0.4222	0.7936	0.4603	**1.0000**
run D – PR	0.4231	0.7941	0.4156	**1.0000**
run E – IN	**0.4383**	0.7937	**0.4851**	**1.0000**
run E – PR	**0.4383**	0.7937	**0.4851**	**1.0000**
run F – IN, IN	0.4258	0.7940	0.4596	**1.0000**
run F – IN, PR	0.4262	**0.7954**	0.4515	**1.0000**
run F – PR, IN	0.4290	0.7941	0.4382	**1.0000**
run F – PR, PR	0.4287	0.7947	0.4320	**1.0000**
baseline	0.4311	0.7859	0.5435	1.0000

7. RE-RANKING USING CO-CITATION INFORMATION

Another popular citation analysis method is to gather information about co-citation patterns. A pair of documents is co-cited when they both occur in the same reference list of a third paper. Citation overlap between documents or frequent co-citation of two documents have proven to be strong indicators for document similarity [30]. Co-citation does not have to be between documents only: two authors can also be co-cited if their papers are both cited by the same third document. Collecting this information for a large group of papers and authors will result in a list that represents the similarity between authors: co-cited authors often tend to write about the same topics. In the experiments described in this section, we investigate using co-citation counts as a means of identifying similar experts. We used these similarities between author to re-rank the expert rankings produced by the approaches of the previous sections.

In early works on citation analysis, co-citation counts of authors were collected only between the first authors of each cited paper, for reasons of computational complexity. In recent years, all-author co-citation analysis is the more frequently used method. The two methods have been frequently compared. We choose to adopt all-author co-citation counts as opposed to first-author co-citations. Because we deal with a small workgroup with only 89 members, using only the first authors would have lead to a much smaller and more incomplete list of co-cited author pairs: only 72.9% of the author pairs would be disregarded then. Furthermore, Schneider et al. (2007) argued that all-author co-citation leads to better distinguishable clusters of authors [31].

Finally, we did not only count authors of different papers occurring together in the reference list as a co-citation pair, but also two authors of the same paper occurring in the reference list. We did not exclude the latter type of co-citation because we regard co-authorship as a measure of author similarity as well. For our test collection and the associated crawled citation network, this counting method resulted in a total of 13,284 unique co-cited author pairs. ILK workgroup members were co-cited with 46.7 other authors on average.

7.1 Co-citation re-ranking

We transform our list of co-citation pairs and the corresponding counts into similarities between authors by normalizing the vector of co-citation counts. We represent the similarity between two authors a_i and a_t as $sim(a_i, a_t)$ and define $sim(a_i, a_i)$ to be equal to 1. Equation 2 represents our re-ranking method using co-citation similarity, which takes into account the similarities with the other authors proportionate to their position on the list.

$$new\_score(a_t, q_k) = \sum_{i=1}^{m} \left(\frac{expertise(q_k, a_i) \cdot sim(a_i, a_t)}{rank(q_k, a_i)} \right) \qquad (2)$$

For each of the m authors in the result list, the influence their expertise score $expertise(q_k, a_i)$ has on the active author[6] a_t is moderated by the similarity between both authors $sim(a_i, a_i)$ as well as by the rank of author a_i on the result list or $rank(q_k, a_i)$. These moderated influences are then summed, forming the new score for the active author $new\_score(a_t, q_k)$. Note that the score of the active author a_t himself is also incorporated in the new score, only moderated by the rank $rank(q_k, a_t)$ since the similarity of the active author with himself is equal to 1. Space restrictions keep us from reporting all the re-ranking methods we experimented with; the described method significantly outperformed all other methods.

We evaluated our methods on the best performing runs from the previous experiments. Table 6 shows the results of re-ranking these

[6] The re-ranked score is calculated for each author separately and each is in turn considered to be the active author.

runs using author co-citation counts, evaluated on the set of 80 queries.

Table 6: Results of re-ranking some of the best-performing runs. Scores printed in bold represent improvements over the original score, evaluated on the set of 80 queries.

	before re-ranking		after re-ranking	
	MAP	**MRR**	**MAP**	**MRR**
run A	0.4311	0.7859	**0.4354**	**0.7877**
run B – PR	0.4262	0.7925	**0.4266**	**0.7927**
run C – PR	0.4341	0.7941	0.4296	0.7939
run D – PR	0.4290	0.7942	**0.4337**	0.7942
run E – PR	0.4383	0.7937	0.4370	0.7876
run F – PR, IN	0.4290	0.7941	**0.4372**	0.7941

The comparison is similar to the comparisons in the previous two sections: although there are some slight improvements in MAP on the 80 queries, applying co-citation analysis does not yield any significant improvements. There were almost no changes in MRR scores after re-ranking.

8. DISCUSSION

In the present study we experimented with two different types of evidence for expertise in an average-sized academic workgroup: the content of academic papers, and the academic-social citation network connecting those papers.

In our relatively small workgroup setting we found no significant differences between using a content-based query-dependent ranking and a static ranking based on the social network. This would imply that—in a workgroup setting—there is no merit in taking into account the topic of the query; the same experts are important for every query. A likely reason for this is that researchers with many citations have co-authored many papers, and have been workgroup members for a longer time. Long-time members have had more time to contribute papers to the workgroup's output and are more likely to have touched multiple topics. This means that they are more likely to have co-authored papers relevant to the original queries. This seems to suggest that simply outputting the most prolific authors would be a good baseline expert finding strategy in a workgroup.

However, preliminary experiments with this only partially confirmed this: a static ranking based on publication count was good at ranking a relevant expert at the top of the list (e.g. a high MRR), but it was not good at finding all the relevant experts, as signaled by the significantly lower MAP scores. This suggest that one of the top three most prolific authors almost always is a relevant expert, but that they are a special case in the workgroup. They tend to skew expert finding results towards them and make it harder to find the other relevant experts.

Another possible reason for the lack of significant differences between pure content-based and pure citation-based approaches could be the narrow focus of a scientific workgroup. In general, the topics of the ILK workgroup focus on machine learning applied to tasks in language technology. This means that it will be harder for the content-based algorithm to distinguish between the different documents (and thus authors) when predicting relevance.

Another conclusion we can draw from our results is that in calculating the importance of network elements, there is no significant difference between using the PageRank algorithm or simply counting the number of citations (the indegree). A possible explanation for this is the relative shallowness of the citation network used in our experiments. With only three layers (the middle layer being the ILK documents, the *citing* and *cited by* layers having been crawled through Google Scholar), calculating PageRank will tend to yield results similar to simply counting the indegree, as there can be only a minor non-uniform distribution of weights among the nodes.

Furthermore, for expert finding within a workgroup setting there does not appear to be a successful way of combining citation analysis with content-based expert finding, as none of our combination yielded any significant improvements. This negative result suggests that the output of the separate approaches overlaps to a large extent, with no measurable complementary power.

Combining one type of citation analysis with another—e.g. indegree on the document citation network with PageRank on the author citation network—also did not yield any real improvements, which suggests that whether PageRank or indegree is used on the author or document citation network, hardly influences the resulting lists. Therefore, the most effective and recommended way of performing expert finding in a workgroup setting seems to be the computationally least intensive method: collecting the indegree of publications. These results are in line with the outcomes of the approaches discussed in section 2.

Evaluating our approaches on the set of 10 queries with real expert relevance judgments resulted in MAP and MRR scores that were higher than for the set of 80 queries in all but two cases. This suggests that our decision to unequivocally equate authorship with expertise did not result in perfect expert relevance judgments. Assumptions such as ours or the one made in the W3C collection should therefore be avoided in the construction of future expert finding test collections. Evaluating on 10 queries is not enough for statistically sound comparisons, but the content-based approach does perform much better than any of the citation-based or combined approaches. However, it is unclear whether this means our conclusions about indegree-based static rankings being the most effective hold when evaluated using a larger set of real expert relevance judgments.

We think that the relatively small size of an average academic workgroup is the main reason for our lack of significant improvements, opening up the question whether this issue could be overcome with other means perhaps. Hypothetically, one option would be to normalize expertise scores by the longevity of the author's workgroup membership. Another way of correcting for the effect of long-time group members, ranked high simply because their credit is accumulated over a longer period, would be to devalue older publications. New publications are fresher in the minds of the workgroup members, so the influence of older papers on the expertise score should be diminished. Further experimentation is needed to investigate the merit of these discounting methods.

Our results and conclusions cannot be generalized to larger contexts, such as a complete research area, a university, or a scientific journal, for which a large-scale computation of author and document networks would be needed.

Despite the negative results on combining methods for expert finding, we do wish to argue based on our results that citation analysis is an effective technique for finding experts in an academic workgroup setting. Using a content-based approach does not add any significant value over straightforwardly counting the number of citations of authors and papers, and ranking the workgroup members on that.

9. FUTURE WORK

Arguably, one way to extend our current experiments would be to test our different approaches on a much larger collection such as the UvT Expert Collection [4], which contains 1,168 university employees divided over many different faculties, departments, and workgroups. The number of topics—1,491 for Dutch and 981 for English—is also an order of magnitude higher. It also contains information about co-taught courses, which would be another social network to investigate. Another advantage of using the complete UvT Expert Collection is the availability of direct expert relevance judgments. Analogous to the W3C collection, our method of producing expert relevance judgments from the original document relevance judgments hinges on the assumption of co-authorship as the proper attribution of expertise.

Another area of improvement of our approach would be to collect a more complete citation network instead of the current two layers around the core document nodes, which might improve the quality of comparison between the PageRank and indegree metrics.

10. ACKNOWLEDGMENTS

We would like to thank Krisztian Balog, Birger Larsen, and Kirstine Wilfred Christensen for fruitful discussions and comments. The work of Toine Bogers is funded by the IOP-MMI-program of SenterNovem / The Dutch Ministry of Economic Affairs, as part of the À Propos project. Antal van den Bosch is funded by NWO, the Netherlands Organization for Scientific Research.

11. REFERENCES

[1] P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The CSIRO Enterprise Search Test Collection. *ACM SIGIR Forum*, 41(2):42–45, December 2007.

[2] K. Balog and M. de Rijke. Determining Expert Profiles (With an Application to Expert Finding). In *Proceedings of IJCAI '07*, pages 2657–2662, 2007.

[3] K. Balog, L. Azzopardi, and M. de Rijke. Formal Models for Expert Finding in Enterprise Corpora. In *Proceedings of SIGIR '06*, pages 43–50, New York, NY, 2006. ACM Press.

[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 SIGIR '07*, pages 551–558, New York, NY, 2007. ACM Press.

[5] S. Bao, H. Duan, Q. Zhou, M. Xiong, Y. Cao, and Y. Yu. Research on Expert Search at Enterprise Track of TREC 2006. In *TREC 2006 Working Notes*, November 2006.

[6] H. Biswas and M. Hasan. Using Publications and Domain Knowledge to Build Research Profiles: An Application in Automatic Reviewer Assignment. In *Proceedings of ICICT '07*, pages 82–86, 2007.

[7] T. Bogers and A. van den Bosch. Authoritative Re-ranking in Fusing Authorship-based Subcollection Search Results. In *Proceedings of DIR 2006*, pages 49–55, Enschede, March 2006. Neslia Paniculata.

[8] T. Bogers and A. van den Bosch. Authoritative Re-ranking of Search Results. In *Proceedings of ECIR 2006*, volume 3936 of *LNCS*, pages 519–522, Berlin, April 2006. Springer Verlag.

[9] T. Bogers, W. Thoonen, and A. van den Bosch. Expertise Classification: Collaborative Classification vs. Automatic Extraction. In *Proceedings of the 17th ASIS&T SIG/CR workshop on Social Classification*, November 2006.

[10] J. Bollen, M. A. Rodriguez, and H. van de Sompel. Journal Status. *Scientometrics*, 69(3):669–687, 2006.

[11] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise Identification using Email Communications. In *Proceedings of CIKM '03*, pages 528–531, New Orleans, LA, 2003.

[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 *TREC 2006 Working Notes*, November 2006.

[13] N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the TREC 2005 Enterprise Track. In *TREC 2005 Working Notes*, November 2005.

[14] S. T. Dumais and J. Nielsen. Automating the Assignment of Submitted Manuscripts to Reviewers. In *Proceedings of SIGIR '92*, pages 233–244, New York, NY, USA, 1992. ACM.

[15] S. Ferilli, N. Di Mauro, T. Basile, F. Esposito, and M. Biba. Automatic Topics Identification for Reviewer Assignment. *Advances in Applied Artificial Intelligence*, pages 721–730, 2006.

[16] A. Fujii. Enhancing Patent Retrieval by Citation Analysis. In *Proceedings of SIGIR '07*, pages 793–794, New York, NY, USA, 2007. ACM Press.

[17] E. Garfield. *Citation Indexing: Its Theory and Application in Science, Technology, and Humanities*. John Wiley & Sons, Inc., New York, NY, USA, 1979.

[18] E. Garfield. Der Impact Faktor und seine richtige Anwendung. *Der Unfallchirurg*, 101(6):413–414, June 1998.

[19] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of DL '98*, pages 89–98, New York, NY, June 1998. ACM Press.

[20] J. E. Hirsch. An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences*, 102(46): 16569–16572, 2005.

[21] P. Jacsó. Google Scholar: the Pros and the Cons. *Online Information Review*, 29(2):208–214, 2005.

[22] J. Kamps and M. de Rijke. The Effectiveness of Combining Information Retrieval Strategies for European Languages. In *Proceedings SAC '04*, pages 1073–1077, 2004.

[23] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, September 1999.

[24] M. Kolla and O. Vechtomova. In Enterprise Search: Methods to Identify Argumentative Discussions and to find Topical Experts. In *TREC 2006 Working Notes*, November 2006.

[25] B. Larsen. *References and Citations in Automatic Indexing and Retrieval Systems - Experiments with the Boomerang Effect*. PhD thesis, Royal School of Library and Information Science, Denmark, 2004.

[26] M. Maybury. Expert Finding Systems. Technical Report MTR 06B000040, MITRE Corporation, 2006.

[27] D. W. McDonald. *Supporting Nuance in Groupware Design: Moving from Naturalistic Expertise Location to Expertise Recommendation*. PhD thesis, University of California, Irvine, 2000.

[28] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[29] G. Pinski and F. Narin. Citation Influence for Journal Aggregates of Scientific Publications: Theory with Application to Literature of Physics. *Inf. Proc. & Man.*, 12(5):297–312, 1976.

[30] G. Salton. Associative Document Retrieval Techniques using Bibliographic Information. *Journal of the ACM*, 10(4):440–457, 1963.

[31] J. Schneider, B. Larsen, and P. Ingwersen. Comparative Study between First and All-Author Co-Citation Analysis Based on Citation Indexes Generated from XML Data. In *Proceedings of ISSI 2007*, pages 696–707, Madrid, 2007.

[32] P. Serdyukov, H. Rode, and D. Hiemstra. University of Twente at the TREC 2007 Enterprise Track: Modeling Relevance Propagation for the Expert Search Task. In *TREC 2007 Working Notes*, November 2007.

[33] I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *TREC 2006 Working Notes*, November 2006.

[34] T. Strohman, W. B. Croft, and D. Jensen. Recommending Citations for Academic Papers. In *Proceedings of SIGIR '07*, pages 705–706, New York, NY, 2007. ACM Press.

[35] D. Yarowsky and R. Florian. Taking the Load off the Conference Chairs: Towards a Digital Paper-Routing Assistant. In *Proceedings of SIGDAT 1999*, pages 220–230, 1999.

[36] J. Zhang, M. S. Ackerman, and L. A. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of WWW '07*, pages 221–230, 2007.

[37] J. Zhu, D. Song, S. Rüger, M. Eisenstadt, and E. Motta. The Open University at TREC 2006 Enterprise Track Expert Search Task. In *TREC 2006 Working Notes*, November 2006.