Design and Evaluation of a University-wide Expert Search Engine

Toine Bogers
Information Interaction and Architecture
Royal School of Library and Information Science
Birketinget 6, DK-2300
Copenhagen S, Denmark
tb@db.dk

Ruud Liebregts
Textkernel BV
Nieuwendammerkade 28A-17
NL-1022 AB
Amsterdam, The Netherlands
liebregts@textkernel.nl

ABSTRACT

We present an account of designing, implementing, and evaluating a university-wide expert search engine. We performed system-based evaluation on multiple query sets to determine the optimal retrieval settings and performed extensive user-based evaluation with three different user groups: scientific researchers, students looking for a thesis supervisor or topic experts, and outside visitors of the website looking for experts. Our search engine significantly outperformed the old search system in terms of effectiveness, efficiency, and user satisfaction.

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

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Expert search, evaluation, search engine, interactive IR

1. INTRODUCTION

The ability to discover individuals that are knowledgeable about a certain topic, task, or assignment is essential for organizational effectiveness and is generally referred to as *expert finding*. Experts can answer questions, point to other specialists, or perform functions that require special knowledge, skill, or experience. Obtaining such a complete and up-to-date overview of "who knows what" in an organization is important for the rapid formation of project teams to respond to new market opportunities or threats [11]. Expert finding is different from *expert profiling*, where the goal is to assess and quantify the range of topics that a person is knowledgeable about [2].

So far, most of the existing expert finding systems have been evaluated and compared under laboratory conditions using static collections, whereas large-scale user evaluations of expert finding systems have been largely absent. In addition, little objective evidence has been presented on the benefits of such dedicated expert search engines or their ability to outperform the existing resources organizations have in place on this task. One notable exception is IBM's Professional Marketplace labor resource management system[1], which provides project managers with the information to efficiently match employee expertise with customer needs, by instantly retrieving data on each of the nearly 180,000 employees' specialties, skills, and location. It is reported to be a success, saving IBM $500 million in its first year, reducing reliance on expensive external contractors, and increasing employee utilization rate.

In this paper we report on the design and large-scale evaluation of an expert search engine for Tilburg University (UvT), a medium-sized university in the Netherlands. The search engine taps into a variety of bilingual sources of topical expertise evidence for over 1,900 university researchers. It presents 'evidence documents' that support the search results and offers additional information, such as the organizational metadata and collaboration statistics of university staff. We built on previous work by [2] and [8] by basing our data on an updated and expanded version of the UvT Expert Collection[2]. In addition, we used real, natural language queries instead of pre-determined topics from an expertise taxonomy for our system-based evaluation. These are both queries generated in TREC-like fashion as well as queries suggested by actual candidate experts. We also experimented with different methods of expertise attribution. Finally, we also performed extensive user-based evaluation of the search engine's performance on two different task types with three different user groups.

The remainder of this paper is organized as follows. In the next section we review related work. Then, in Section 3 we describe the data sources covered by our search engine and our design choices. We performed an extensive evaluation of our expert search engine, which is described in Sections 4–6. Section 4 covers the system-based evaluation, with Section 5

DIR 2010 January 25, 2010, Nijmegen, the Netherlands.

[1] http://www-01.ibm.com/software/success/cssdb.nsf/CS/CSAL-6VQSUU
[2] Available at http://ilk.uvt.nl/uvt-expert-collection/

detailing the evaluation phase using real university experts. Section 6 then describes our evaluation using different tasks and user groups. We conclude our paper with a discussion of our results and goals for future work in Sections 7 and 8.

2. RELATED WORK

Early approaches to expert finding, such as the Plexus system, an expert system on gardening designed as a referral tool for public libraries [16]. It required people to create their own profile by describing their field(s) of expertise, or to assess their skills against a predefined set of keywords. The disadvantage of such people-finders is that they place an unnatural workload on employees and system administrators to keep the system up-to-date. In addition, the resulting expertise descriptions are usually incomplete or too general. This prompted a shift to expert finding techniques more supportive of the natural expertise location process, such as Campbell et al., who developed a content-based approach to expert finding by mining e-mail collections [5]. Using the social networks present in the e-mail traffic, they were able to outperform a purely content-based approach. Expert finding received its biggest boost when it was included in one of the 2005 TREC tracks. Many different retrieval models for expert finding have been proposed and evaluated by the TREC participants [14].

In 2007, Balog et al. released the UvT Expert Collection, which is based on a bilingual (Dutch and English) database of university employees who are involved in a broad range of research or teaching areas [2]. Four document types were extracted from this database: research descriptions, course descriptions, publication metadata (and full text where available), and academic home pages. In contrast to the collections used in TREC, document-candidate associations are clear and the data is structured and clean, albeit sparse. The collection also contains information about the organizational structure. Balog et al. found that expert finding models developed for TREC collections generalized well to this new setting. Their experiments also revealed that expert finding in such a knowledge-intensive setting benefits most from publication-based evidence, whereas academic home pages contributed least to expert finding.

Not many studies have examined user interaction with expert finding systems. Shami et al. found that result order and social connection are the main factors that influence the user's decision to further explore an expert search result [13]. While the result order is also an important factor in Web search, the role of social connections between the user and the presented results was found to be significant only in expert search selections.

3. DESIGN AND IMPLEMENTATION

3.1 Data Sources

We used the university address book, accessible through the LDAP protocol, as our list of candidate experts. We extracted 1,944 potential experts and their metadata from the address book and stored this information, including a unique administrative identifier (ANR), in a MySQL database. This is a larger number of potential experts than contained in the UvT Expert Collection, because of the growth of the university.

Tilburg University currently has a number of systems available that can be used to support expert finding. An important source of expertise evidence are the document repositories that contain over 40,000 scientific publications and over 12,500 student theses. Users can search through the metadata fields for relevant publications or theses, but full text search is not available. Using these document repositories to bridge the gap between topics and the appropriate experts hinges on the assumption that paper authors and thesis supervisors have expertise about the document topics. However, this requires users to manually search for papers and theses and associate them with people. Moreover, the authors and supervisors displayed may no longer be working for the university. In addition to these two sources, the university allows researchers to maintain their own profile on Webwijs ("Webwise")[3], an online database of university experts and expertise. Researchers can enter research descriptions and select expertise areas from a predefined list. It also links to courses taught by the expert and authored publications. Users can search for experts on Webwijs by expert name—unknown in an expert finding scenario—or by topic. Searching by topic has the disadvantage that the Webwijs system only accepts queries that match one of the predefined expertise key words, reducing flexibility. Finally, the fourth source of information is the search engine for the UvT intranet, which is based on ht://Dig[4]. The retrieved Web pages are the most difficult of all three sources to associate with candidate experts.

3.2 Index construction

The repository of publications and theses was queried using the OAI-PMH protocol[5]. We extracted all English and Dutch documents from the publication and thesis repositories, indexing all the relevant metadata, such as title, author(s), and publication date, and the full text when available. The metadata was also stored in our database to enable fast retrieval for user interface purposes and generating association statistics such as paper collaboration. People were unambiguously associated with publications using the ANRs of the authors. ANRs of thesis supervisors were not available in the repository, so we had to use pattern matching techniques to match the metadata names to candidate experts, resulting in a small percentage of possible association errors. We additionally crawled the Webwijs page of each candidate expert, if available. All information from Webwijs—research descriptions, expertise areas, and course descriptions—were stored in a single Webwijs profile document for each candidate expert. Academic home pages were not included because of their marginal contribution to expert finding [2].

The generated XML documents were indexed using Indri 2.6[6]. English and Dutch stop words were removed and English Krovetz stemming was applied to all documents, including those written in Dutch, because Indri provides no Dutch stemming algorithm. All queries were stemmed the same way. We do not believe this decision about stemming influenced the validity of our results. The documents and metadata extracted from Webwijs and the UvT repositories can easily be updated and re-indexed, ensuring an up-to-date expert search engine.

[3] http://www.tilburguniversity.nl/webwijs/
[4] http://www.htdig.org
[5] http://www.openarchives.org/pmh/
[6] http://www.lemurproject.org/indri/

3.3 Expert finding

With so many fragmented sources of expertise evidence as in the UvT situation, it is difficult for users to obtain a coherent picture. Our expert search engine therefore combines the content-based evidence available on Webwijs and from the repositories. In our search engine, we opted for a document-centric approach to expert finding. Previous work has shown that a document-centric approach to expert finding works well and that is a robust model for expert finding [1].

A document-centric approach consists of three steps, as illustrated in Figure 1. First, normal *document retrieval* is used to match the query to relevant documents in the collection. For document retrieval, we used the Indri toolkit, which combines the benefits of Bayesian inference networks and a statistical language modeling framework [15]. Indri also supports a dependence model based on term proximity as well as pseudo-relevance feedback. Preliminary experiments suggested using Dirichlet smoothing.

In the next step, *expert association*, the retrieved documents are associated with the candidate experts. In our case, this step was performed at indexing time and could be done unambiguously in the large majority of cases. The associations are then looked up again at retrieval time. *Expertise attribution* is the third and final step, where a (normalized) relevance score is assigned to each candidate expert and the top N relevant experts are presented to the user. This step is necessary, because more than one document in the result set may be associated with the same expert, and single documents might be associated with multiple candidates. Expertise attribution can be done in different ways, where one of the most straightforward scoring methods simply involves counting the number of query-relevant documents associated with each expert. However, this ignores the score and rank of the retrieved documents. Another simple method would be to take the score of the first document in the result set that is associated with an expert. However, this approach ignores the presence of multiple relevant documents per expert, which may also be an indicator of expertise. Balog et al. computed expertise scores by simply summing the relevance scores of the documents associated with an expert [1]. When many highly relevant, retrieved documents are associated with a candidate, his or her expertise score will be high. Bogers et al. introduced a method for expertise attribution that discounts for document ranks [3]. It further moderates the rank's influence by taking the logarithm of the rank, as shown in Equation 1

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

where for all m documents authored by author a_t the relevance scores $score(q_k, d_i)$ are divided by the logarithm of the rank $rank(q_k, d_i)$ plus 1 (to avoid division by zero), and then summed to form a single expertise score for that author-query pair. As a result of this discounting, documents further down the result list are penalized less. The logarithm, however, is still a strong discounting factor and we therefore experimented with methods that punish the lower-ranked documents less. We ended up with a weighted combination of the rank reciprocal of a document and the document's relevance score on that query. Equation 2 shows this expertise attribution scheme.

$$expertise(a_t, q_k) = \sum_{i=1}^{m} \left(score(q_k, d_i) + \frac{2}{rank(q_k, d_i) + 1} \right) \quad (2)$$

Again, this leads to documents further down the result list being assigned a lower score. This method was originally parameterized but we only report the formula with optimal parameter values here due to space restrictions.

Finally, in addition to expertise attribution based on score and rank of retrieved documents, data fusion—combining multiple result sets into one—can also help creating an expertise score, according to, among others, [7] and [10]. In addition, static rankings or prior probabilities can be assigned to documents or candidate experts and used to re-rank the search results. We refer the reader to the work mentioned in Section 2 for some examples of such re-ranking methods, such as [5].

4. SYSTEM-BASED EVALUATION

The goal of our system-based evaluation was to evaluate the different options available during the design stage in a laboratory setting. We can then use the best-performing approach to expert finding for our user-based experiments. To perform such an evaluation under laboratory conditions, without participating users, we needed a set of representative queries and the appropriate relevance judgments at the expert level. We set out to construct such a test set using the known associations between documents and experts: we assumed that the document authors or supervisors were the relevant experts for topical queries derived from those documents. We randomly selected 120 publications and 120 theses and divided these equally over the English and Dutch documents in our collection. For each of these random documents, a short query topic was derived from the title and abstract of each document, with average query length being 2.1 words. This resulted in a set of 240 queries, 120 for each language, each subset spanning 60 documents and 60 theses.

Relevance judgments were assumed to be binary and the relevant experts for each query were the associated authors or supervisors. There are some drawbacks to this way of generating relevance judgments. Normally, relevance judgments are elicited from users or independent judges, not from simple document-candidate associations. Furthermore, the relevance judgments do not cover candidates that are not associated with the test document, but do have expertise on the topic. The judgments may also be unreliable because they are based on two imperfect assumptions: (1) an author of a publication might not have contributed to the actual contents, and (2) a thesis supervisor is not necessarily an expert on the thesis subject. Furthermore, our own assessment of topics derived from the documents may be erroneous. However, Carterette et al. showed that using a large amount of queries should mitigate these disadvantages to some extent [6].

The documents from which the queries were derived, were removed from the result lists for those queries, since querying the index for a topic that was inferred from a document in the index is a form of known-item search. Results would be biased because the test document is very likely to be retrieved.

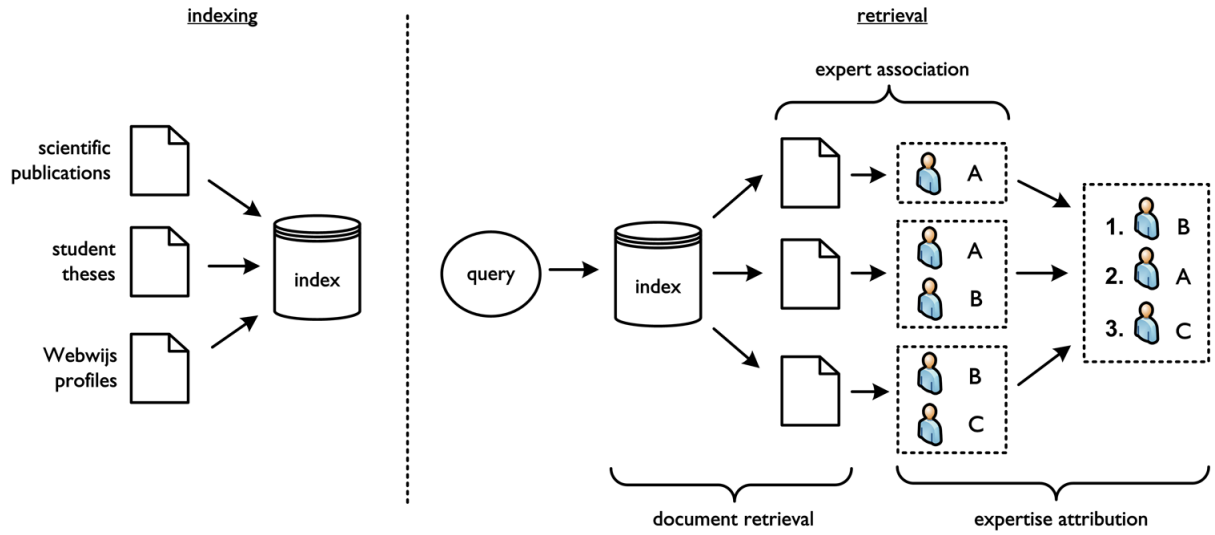

Figure 1: The document-centric approach to expert finding.

4.1 Methodology

We investigated a number of different options for the design of our search engine. We first experimented with the impact of using the dependence model and query expansion based on pseudo-relevance feedback. Expertise attribution was kept simple and constant by using only the score of the first document associated with each expert. Next, we kept the document retrieval side constant and tested the different expertise attribution methods discussed in Section 3.3. We also performed experiments with data fusion by creating different indexes for each information source and then combining the result lists from those retrieval runs.

System performance was measured using Mean Average Precision (MAP) and average Normalized Discounted Cumulated Gain (NDCG) at rank 10. If a query returned less than 10 results, NDCG at the highest rank was taken instead. Because the majority of the documents in our collection are associated with only a single expert, there is usually only one relevant expert per query. As a result, MAP correlates strongly with Mean Reciprocal Rank (MRR). We therefore did not include MRR in our system-based evaluation. To test for statistical significant improvements, we used two-tailed, matched pairs Student's t-tests and looked for improvements at a 0.95 confidence level.

4.2 Results

Table 1 contains the most important results of our system-based evaluation. Our standard language modeling baseline achieved a MAP score of 0.4712 and a NDCG of 0.6018. Both the dependence model and the pseudo-relevance feedback significantly improved on the baseline model, with the dependence model working best on English queries (15–19% improvement), and query expansion working best on Dutch queries (11–17% improvement). The combination was found to perform even better at a MAP of 0.5849 and a NDCG of 0.6926. We therefore selected this for our baseline document retrieval settings. Query expansion slowed down query execution time, but not to unacceptable values (0.5 second per query).

Expertise attribution results are shown in the bottom half of Table 1. The three methods described in Section 3.3 all improved significantly on the baseline. The scheme of weighting the document scores and ranks also outperformed the other two approaches significantly, generating 12–15% improvement over our expertise attribution baseline. Equation **??** is the second best performer, but still performs significantly worse than Equation 2.

We tested the influence of each of the three data sources separately—publications, theses, and expertise profiles—and found that no index containing just one source exceeded the performance of a full index. Publications made the largest contribution to the results, followed by student theses and expertise profiles. As for our data fusion experiments: we tried all of the different fusion methods mentioned in [10]. However, all methods performed significantly worse than when a single index containing all data was used.

5. EXPERT-BASED EVALUATION

After determining the optimal settings of our expert search engine in a laboratory setting, we created a prototype and we performed our first evaluation of it with real users. We enlisted the help of 30 UvT researchers, selected proportionately from the different faculties, to create a new test set with realistic topics and reliable relevance judgments.

5.1 Methodology

Before exposing participants to the expert search engine, we asked each of them to write down one topic about which they themselves were knowledgeable. We also asked them to name up to five other experts on that topic, also employed by the UvT. Since co-workers are relatively good at making judgements about each other's expertise [12], participants were asked to rate their own and their colleagues' relative expertise level on a five-point scale (from 0 = 'no expertise' to 4 = 'high expertise'). Participants were then shown the search engine and asked to use it to find experts on their pre-specified topic. Query formulations were allowed until the participants were satisfied with the results. For the final query and result list, participants were asked to judge each of the top 10 candidates returned by the search engine, providing the candidate was known to them. We concluded with a short survey to measure satisfaction on a

Table 1: Results of our system-based evaluation. The top half of the table shows results of the different document retrieval options; the bottom half shows the results of the different expertise attribution methods. Best scores for each half are printed in bold.

	System-based query set (240 queries)			
Method	**MAP**	**% change**	**NDCG@10**	**% change**
Language modeling baseline	0.4712		0.6018	
Dependence model	0.5185	+10.0%	0.6527	+8.5%
Query expansion	0.5308	+12.6%	0.6398	+6.3%
Dep. model & query expansion	**0.5849**	+24.1%	**0.6926**	+15.1%
First doc (baseline)	0.5849		0.6926	
Score sum	0.6529	+11.6%	0.7654	+10.6%
(Eq. 1) Log rank discount	0.6326	+8.1%	0.7518	+8.7%
(Eq. 2) Weighted rank-score comb.	**0.6757**	+15.5%	**0.7755**	+12.1%

five-point Likert scale (from 1 = 'strongly disagree' to 5 = 'strongly agree'). Participants were observed individually during these sessions. This resulted in an 'expert-centered' set of 30 new queries with 268 graded and realistic relevance judgments, enabling us to verify our findings from the system-based evaluation. Because the expert-centered test set contains up-to-date topics, we were also able to evaluate the 'up-to-dateness' and 'position' static rankings [8]. Furthermore, we tested whether the number of times a document or candidate had been cited could be used to re-rank the results. We used Google Scholar as a source of citation counts.

5.2 Results

Out of the 30 participants, 18 never reformulated their initial query and most of the reformulations of the other 12 were only translations of the original queries, which is also reflected in the average query length of 1.6 terms. The 30 new topics came with 268 graded relevance judgments, making NDCG a more important measure for evaluation. Running this query set, the highest MAP and NDCG values we obtained were 0.7982 and 0.8071 respectively. Re-testing our original design options using this new query set, we no longer found a significant difference between the three main expertise attribution methods using this test set. Data fusion methods or (combinations of) single indexes still did not outperform using a single index containing all three sources. Finally, the survey results showed a high degree of satisfaction with the search results with an average score of 3.77 (SD = 0.90). Participants were also highly satisfied with the search engine's speed (M = 4.33; SD = 0.88), the evidence shown to support the results (M = 3.53; SD = 1.20), and the up-to-dateness of the results (M = 3.80; SD = 1.13). There were also some constructive comments that helped us further improve the search engine. For instance, the theses were perceived to "pollute" the search results in some cases. This issue was addressed by creating selection boxes for each of the three source types, allowing the user to include or exclude them.

Tests with static rankings showed that the up-to-dateness factor improved the results slightly, but the improvement was not statistically significant. The other static rankings (position and citation counts) degraded performance slightly. Because Google Scholar covered only 30% of the publications in our collection, the prior probabilities we created were based on incomplete information. However, the setting that [3] used, where information was more complete (around 80% coverage), did not render any significant improvement either.

Because some participants suggested that student theses "polluted" the results, we tested all possible combinations of data sources and their contribution to system performance. Using all data sources still produced the best results, for which publications are the largest contributor. Although expert profiles alone have an average contribution, omitting them from the index resulted in only a 12.15% decrease in performance. Removing student theses resulted in a nearly 18% performance loss.

6. USER-BASED EVALUATION

In our third and final evaluation step we involved real users to benchmark the performance of the expert search engine ('new system') against all other information sources currently available within Tilburg University combined ('old system'). Section 3 contains an overview of the sources currently available. We used two different groups of participants: an *internal* group of UvT students who had prior knowledge of the current systems and may have been familiar with some of the topics and experts, and an *external* group consisting of high school students who had no such prior knowledge.

6.1 Methodology

In this experiment, users were given two different types of simulated work tasks—semantically open descriptions of the scenario and context of a work task [4]. Our two task types, *expert finding* and *supervisor finding*, included a description of the topic and an indicative request that expressed the task type. Supervisor finding is essentially a special case of expert finding: users will be looking for potential thesis supervisors within a certain organizational unit. For each task and topic pair, users were asked to specify a top three of experts (or supervisors) on a given topic.

We used the topic set from the expert-based evaluation to construct five work tasks of each type and used the associated relevance judgments to evaluate the selections participants made. In addition, we performed a manual relevance assessment of the candidate experts that were recommended frequently by our participants, but were not covered by the expert-based relevance judgements. We are aware that we undercut the expert's judgments here, as our own assessment is unlikely to be 100% correct[7]. In order to reduce

[7]We therefore validated all evaluations using only the orig-

the variation in search strategies used by participants, we used a within-subjects design, in which all participants used both system types an equal number of times (if possible). Display order of the tasks was random, and task types and systems to be used were also assigned randomly. All participants were required to complete at least four tasks and at most ten, and they were asked to select up to three relevant experts.

The experiment was entirely Web-based so users could participate from any location. Figure 3 shows a screenshot of the interface. The tasks and the expert selections were shown in a pane on the left side of the screen, as well as the available sources for each current task. Selecting a source—the new system or any of the old systems—opened the search engine interfaces in a large frame to the right of the task pane, so that both were visible simultaneously. At the end of the experiment, we conducted a short survey with questions that together covered all aspects of usability: effectiveness, efficiency, and satisfaction. Again, this was measured on a five-point Likert scale as in Section 5.1.

6.2 Evaluation

To benchmark system performance, we used the *immediate accuracy* and *qualified search speed* metrics introduced by Käki. An advantage of these measures is that they are proportional and can therefore be used in cross-system comparisons [9]. *Immediate accuracy* (IA) measures effectiveness and is expressed as a success ratio. It is defined as the percentage of the cases where at least one relevant answer was selected by the user after making k selections. For example, an IA of 85% after two selections means that in 85% of the cases at least one of the top two selected results were relevant. *Qualified search speed* (QSS) trades off speed against accuracy by including the answer quality based on the graded relevance judgments. It is therefore a measure that captures both effectiveness and efficiency. QSS is expressed as the number of answers found in a particular relevance category divided by the time spent searching.

6.3 Results

A total of 101 users participated in the experiment, of which 44 were part of the external group and 57 were internal participants, selected equally from all faculties. Average age in the internal group was 22.23 (SD = 3.21) and 16.95 (SD = 0.65) in the external group. All participants reported to be familiar with Web search engines. A total number of 325 tasks were completed using the new expert search engine and 332 using the old system. Figure 2 shows the results of the comparison between the two systems.

Our expert search engine was found to be significantly more effective than the old system, as measured by immediate accuracy. Users were able to recommend at least one relevant expert after three selections in 73.6% of the cases when using the old system versus 94.5% with the expert search engine. The main cause for this difference is the poor performance of external participants when using the old system. They performed significantly worse on both task types and achieved an IA of only 68.8%, while the internal group found at least one good answer using the old system in 85.8% of the cases. In contrast, differences in IA between the internal and the external groups using the new system are statistically insignificant for both task types. This suggests that users with no prior experience with the old system are at a disadvantage. In fact, the externals even performed slightly better than the internal group when using the expert search engine, which suggests there is almost no learning curve for the expert search engine.

Evaluation using QSS showed that the expert search engine is both more effective and efficient. For the expert search engine, the majority of the relevant answers (65.8%) are in high relevance categories ($\geq$3), versus 44.3% for the old system. The expert search system also has more relevant answers in the highest category than the old system (38.7% vs. 21.0%), while the old system has more irrelevant answers (36.3% vs. 15.7%). The new system not only finds more highly relevant answers than the old system, it also generates over three times as many highly relevant answers per minute (0.71 vs. 0.21). The external group benefits most from using the expert search engine: QSS of highly relevant answers increased from 0.19 to 0.87 (over 4 times faster). Internal users found those answers 2.6 times faster (0.22 vs. 0.58 answers per minute), once again illustrating the absence of a learning curve of the new system. Finally, users also showed high satisfaction with all aspects of the search engine.

7. DISCUSSION & CONCLUSIONS

In this paper, we presented the design and evaluation of a university-wide expert search engine, that uses content-based evidence in the form of publications, supervised theses, and expertise profiles to generate expert search results. We performed system-based evaluation of our document-centric approach on multiple query sets to determine the optimal retrieval settings. In addition, we performed an extensive user-based evaluation with 3 different user groups: scientific researchers, students looking for a thesis supervisor or topic expert, and outside visitors of the website looking for experts.

Our results show that an integrated approach to expert finding, such as our expert search engine offers, yields several benefits. Our search engine scores significantly better on user satisfaction, efficiency, and effectiveness when benchmarked against the current systems available at the university. Users find more highly relevant answers using our search engine, and find them significantly faster than in the old situation. Perhaps most importantly, we show that our integrated approach has no learning curve for outside users: inexperienced users are able to use the search engine just as successfully and effectively as users already familiar with the university.

8. FUTURE WORK

There are several promising avenues for future work. One we are currently investigating is how different interfaces influence the effectiveness, efficiency and satisfaction of the search engine. Are visual representations of the social network useful, and do 'evidence' documents or snippets increase satisfaction? Query log analysis is another interesting option: analyzing the query logs after a longer period of university-wide usage of the search engine might yield interesting results. Finally, there is more content-based expertise evidence available, such as press releases and project Web pages. According to [8], media experience can be an impor-

inal relevance judgements as well, but found no significant differences between the two sets.

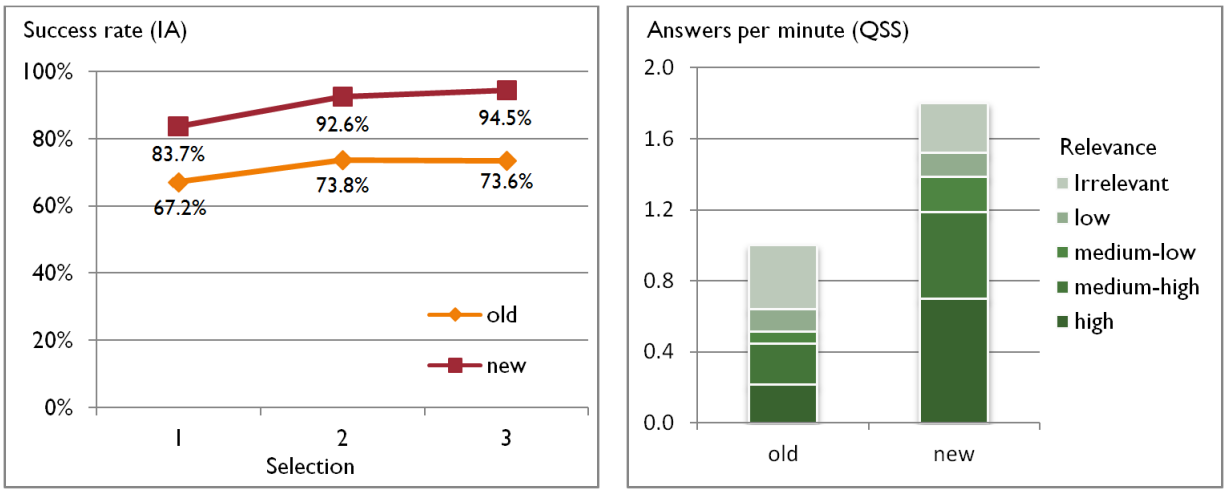

Figure 2: Results of the user-based evaluation of both systems. Figure 2(b) shows the immediate accuracy (or success rate) after k selections, while Figure 2(a) shows the qualified search speed of the two systems.

tant indicator of expertise, so this might be used to improve the search result rankings further.

9. REFERENCES

[1] 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, 2006. ACM.

[2] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad Expertise Retrieval in Sparse Data Environments. In C. L. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, editors, *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 551–558, New York, NY, July 2007. ACM.

[3] T. Bogers, K. Kox, and A. Van den Bosch. Using Citation Analysis for Finding Experts in Workgroups. In E. Hoenkamp, M. de Cock, and V. Hoste, editors, *Proceedings of the 8th Belgian-Dutch Information Retrieval Workshop (DIR 2008)*, pages 21–28, April 2008.

[4] P. Borlund. Experimental Components for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation*, 56(1):71–90, February 2000.

[5] 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 Orleans, LA, 2003.

[6] B. Carterette, J. Allan, and R. Sitaraman. Minimal Test Collections for Retrieval Evaluation. In *SIGIR '06: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268–275, New York, NY, USA, 2006. ACM.

[7] E. A. Fox and J. A. Shaw. Combination of Multiple Searches. In *TREC-2 Working Notes*, pages 243–252, 1994.

[8] K. Hofmann, K. Balog, T. Bogers, and M. de Rijke. Integrating Contextual Factors into Topic-centric Retrieval Models for Finding Similar Experts. In *Proceedings of ACM SIGIR 2008 Workshop on Future Challenges in Expert Retrieval*, pages 29–36, July 2008.

[9] M. Käki. Proportional Search Interface Usability Measures. In *NordiCHI '04: Proceedings of the Third Nordic Conference on Human-Computer Interaction*, pages 365–372, New York, NY, USA, 2004. ACM.

[10] J. Kamps and M. De Rijke. The Effectiveness of Combining Information Retrieval Strategies for European Languages. In *Proceedings 19th Annual ACM Symposium on Applied Computing*, pages 1073–1077, 2004.

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

[12] D. W. McDonald. Evaluating Expertise Recommendations. In *Proceedings of the ACM 2001 International Conference on Supporting Group Work (GROUP'01)*, pages 214–223, 2001.

[13] N. S. Shami, K. Ehrlich, and D. R. Millen. Pick me!: Link Selection in Expertise Search Results. In *CHI '08: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1089–1092, New York, NY, USA, 2008. ACM.

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

[15] T. Strohman, D. Metzler, and W. B. Croft. Indri: A Language Model-based Search Engine for Complex Queries. In *Proceedings of the International Conference on Intelligence Analysis*, May 2005.

[16] A. Vickery, H. Brooks, and B. Vickery. An Expert System for Referral: The PLEXUS Project. *Intelligent Information Systems: Progress and Prospects*, pages 154–183, 1986.

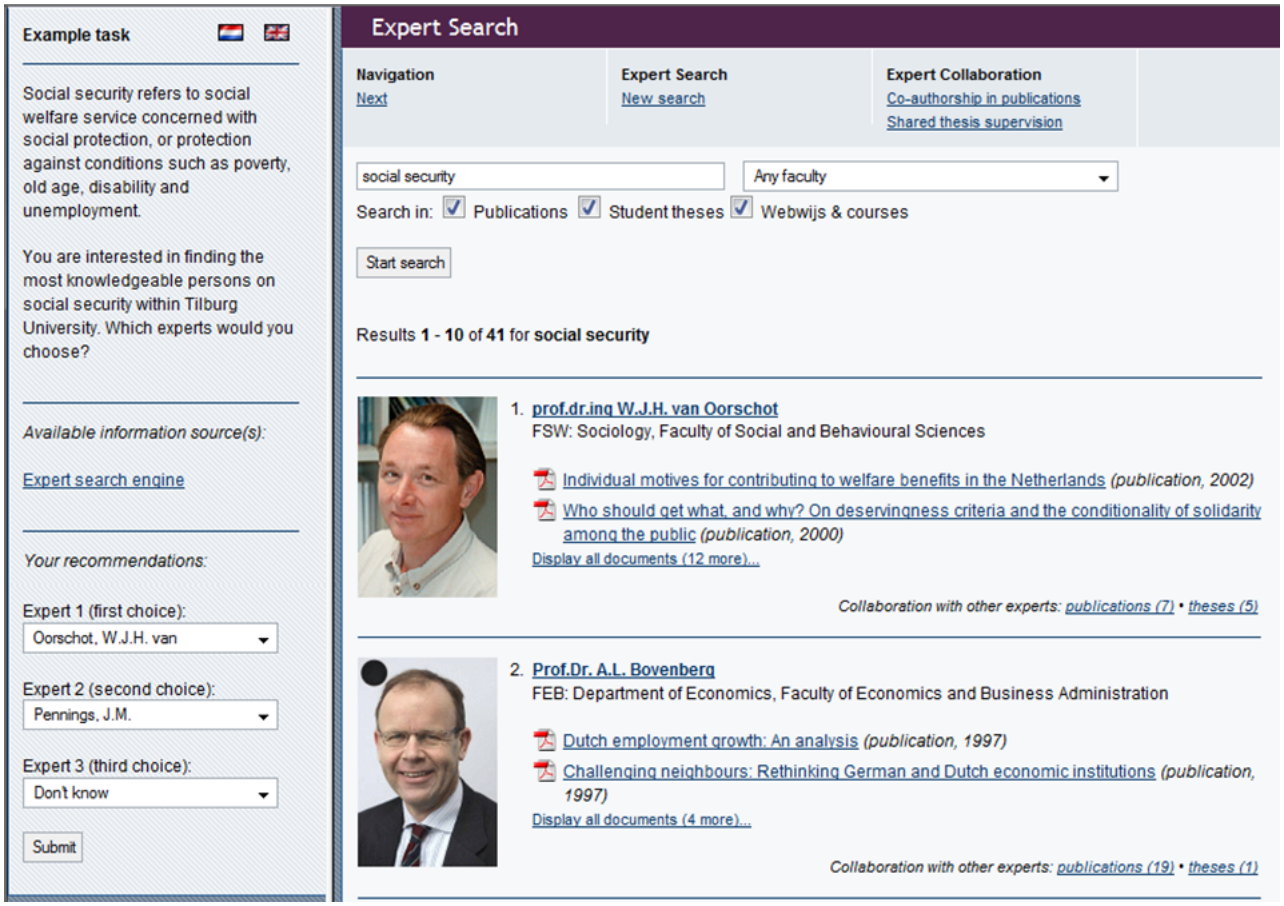

Figure 3: Screenshot of the interface for the user-based evaluation. Tasks and expert selection are displayed in the left pane, and the search engine interfaces in the right pane. Our expert search engine is currently displayed in the right pane.