Integrating contextual factors into topic-centric retrieval models for finding similar experts

Hofmann, K.; Balog, K.; Bogers, T.; de Rijke, M.

Publication date
2008
Document Version
Final published version
Published in
fCHER: SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval: Proceedings

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: $D(e)$	$\frac{\|D(e) \cap D(f)\|}{\|D(e) \cup D(f)\|}$
TERMS	documents	vector: $\vec{t}(e)$	$\cos(\vec{t}(e), \vec{t}(f))$
TOPICS	expertise areas	set: $T(e)$	$\frac{\|T(e) \cap T(f)\|}{\|T(e) \cup T(f)\|}$

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.

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

[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.

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*, *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.

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

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*

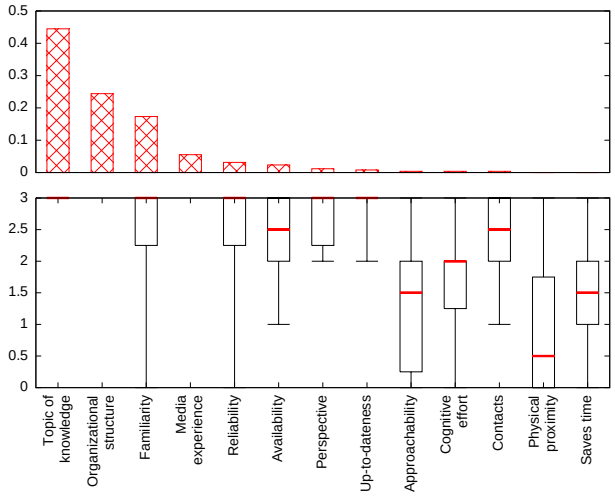

Figure 1: Frequency of implicit factor mentions (above) versus explicit ratings (below). For explicit ratings median, quartiles, minimum and maximum ratings are indicated. For *organizational structure* and *media experience* no explicit ratings are available as these factors were only identified during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