# Contextual Factors for Finding Similar Experts[1]

**Katja Hofmann and Krisztian Balog**
*ISLA, University of Amsterdam, Science Park 107, Amsterdam, The Netherlands.*
*E-mail: {K.Hofmann, K.Balog}@uva.nl*

**Toine Bogers**
*ILK, Tilburg University, P.O. Box 90153, Tilburg, The Netherlands. E-mail: A.M.Bogers@uvt.nl*

**Maarten de Rijke**
*ISLA, University of Amsterdam, Science Park 107, Amsterdam, The Netherlands.*
*E-mail: mdr@science.uva.nl*

**Expertise-seeking research studies how people search for expertise and choose whom to contact in the context of a specific task. An important outcome are models that identify factors that influence expert finding. Expertise retrieval addresses the same problem, expert finding, but from a system-centered perspective. The main focus has been on developing content-based algorithms similar to document search. These algorithms identify matching experts primarily on the basis of the textual content of documents with which experts are associated. Other factors, such as the ones identified by expertise-seeking models, are rarely taken into account. In this article, we extend content-based expert-finding approaches with contextual factors that have been found to influence human expert finding. We focus on a task of science communicators in a knowledge-intensive environment, the task of *finding similar experts*, given an example expert. Our approach combines expertise-seeking and retrieval research. First, we conduct a user study to identify contextual factors that may play a role in the studied task and environment. Then, we design expert retrieval models to capture these factors. We combine these with content-based retrieval models and evaluate them in a retrieval experiment. Our main finding is that while content-based features are the most important, human participants also take contextual factors into account, such as media experience and organizational structure. We develop two principled ways of modeling the identified factors and integrate them with content-based retrieval models. Our experiments show that models combining content-based and contextual factors can significantly outperform existing content-based models.**

Received March 16, 2009; revised November 19, 2009; accepted November 19, 2009

[1]This is an expanded and revised version of Hofmann, Balog, Bogers, & de Rijke, 2008.

 DOI: 10.1002/asi.21292

## Introduction

The increasing amount of information available is making the need to criticially assess information more important. The burden of credibility assessment and quality control is partly shifting onto individual information seekers, but the need for information intermediaries (e.g., experts) has not disappeared and is actually increasing in cases where the credibility of information has to meet high standards (Metzger, 2007). Against this background, *expert finding* is a particularly relevant task: identifying and selecting individuals with specific expertise, for example, to help with a task or solve a problem. Expert finding has been addressed from different viewpoints, including expertise retrieval, which takes a mostly system-centered approach, and expertise seeking, which studies related human aspects.

The goal of *expertise retrieval* is to support search for experts using information-retrieval technology. Following the experimental paradigm and evaluation framework established in the information-retrieval community, expertise retrieval has been addressed in world-wide evaluation efforts (Craswell, de Vries, & Soboroff, 2006). Promising results have been achieved, particularly in the form of algorithms and test collections (Bailey, Craswell, Soboroff, & de Vries, 2007; Balog, 2008). State-of-the-art retrieval algorithms model experts on the basis of the documents with which they are associated, and retrieve experts on a given topic using methods based on document retrieval, such as language modeling (Balog, Azzopardi, & de Rijke, 2009; Balog, Soboroff, et al., 2009). In evaluations of these algorithms, user aspects have been abstracted away.

While research into expertise retrieval has primarily focused on identifying good topical matches between needs for expertise and the content of documents associated with candidate experts, behavioral studies of human *expertise*

*seeking* have found that there may be important additional factors that influence how people locate and select experts (Woudstra & Van den Hooff, 2008); such factors include accessibility, reliability, physical proximity, and up-to-dateness. We term these *contextual factors* to distinguish them from content-based factors that have been explored in previous work (discussed later).

*Context* can be considered to encompass many dimensions, including factors relating to the organization, information seeker, information objects, work task, search task, and so on (cf. Cool & Spink, 2002; Ingwersen & Järvelin, 2005; Kelly, 2006), but here we focus on one specific dimension: contextual factors related to the information objects, in our case—experts. Other elements may play a role, but their impact cannot be studied in our setup: task, organization, and information seeking system are fixed; individual differences between information seekers are left out of our consideration.

Our aim in this article is to explore the integration of contextual factors into content-based retrieval algorithms for finding similar experts. We look at this problem in the setting 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: The top expert identified by a communication advisor in response to a request is not available because of meetings, vacations, sabbaticals, or other reasons. In this case, communication advisors have to recommend similar experts, and this is the setting for our expert finding task. Based on this task, we address three main research questions:

- Which contextual factors influence (human) decisions when finding similar experts in the university setting we study?
- How can such factors be integrated into content-based algorithms for finding similar experts?
- Can integrating contextual factors with existing, content-based approaches improve retrieval performance?

To answer our research questions, we proceed as follows. Through a set of questionnaires completed by a university's communication advisors, we identify contextual factors that play a role in how similar experts are identified in this situation, and we construct a test dataset to evaluate retrieval performance. We evaluate both content-based approaches and approaches where we integrate contextual factors.

The contribution of the article is threefold. For a specific expert-finding task, we succeed at identifying factors that humans use to select similar experts. We model several identified factors and integrate these with existing, content-based approaches to finding similar experts. We show that our new models can significantly outperform previous approaches. Our results demonstrate that it is possible to identify and model contextual factors in the studied task of finding similar experts, and we think that this may be the case for other retrieval tasks as well.

The remainder of the article is organized as follows. We first provide background information on human expertise seeking and expertise retrieval. Next, we describe our approach, including the methods used for data collection, retrieval models and retrieval evaluation. Our results are then given and discussed. Lastly, we present conclusions and outline future work.

## Background

Research on how to enable people to effectively share expertise can be traced back to at least the 1960s when studies in library and information science explored what sources of information knowledge workers such as researchers and engineers use (Menzel, 1960). Subsequent work has identified complex information-seeking strategies relying on a variety of information sources, including human experts (Hertzum, 2000; Rosenberg, 1967).

From results of this research (and other influences) grew the realization that the expertise of employees is a major value of an organization and that effective sharing of knowledge can lead to material gains (Davenport & Prusak, 1998; Sproull & Kiesler, 1996; Wiig, 1997). The field of knowledge management developed, with the goal of using knowledge within an organization as well as possible. One focus was on developing information systems that could support search for expertise. Initial approaches were mainly focused on how to unify disparate and dissimilar databases of the organization into a single data warehouse that could easily be mined (ECSCW'99 Workshop, 1999; Seid & Kobsa, 2000). Resulting tools rely on people to self-assess their skills against a predefined set of keywords, and often employ heuristics generated manually based on current working practice.

Despite the achievements made so far, the question of how to provide effective access to expertise is far from solved, and continues to be addressed from different viewpoints. In human-centered research, which we term *expertise seeking*, one of the goals has been to develop descriptive and prescriptive models of how human information sources are used. Some of this work forms the basis for our study in that it helps us to identify the factors that play a role in finding similar experts in the setting we study. We present a selection of work relevant to this article in the next section.

In system-centered work, for which we use the term *expertise retrieval*, one focus has been on the development of effective retrieval algorithms. This work forms the basis for the retrieval aspects of our study, including the content-based baseline retrieval algorithm, relevance assessment, and evaluation methodology. An overview of recent work in this area is given later.

## Expertise Seeking

Of the human-centered research that we term *expertise seeking*, we are particularly interested in models of how people choose an expert. Most relevant are models that identify specific factors that may play a role. We need to be able to identify quite specific factors so that we will be able to model these and integrate them into a retrieval model.

Several studies have identified factors that may play a role in decisions of what expert to contact or recommend.


DOI: 10.1002/asi

TABLE 1. Factors found to influence expert selection by Woudstra and Van den Hooff (2008).

| Factor | Description |
|---|---|
| Quality-related factors | |
| 1. Topic of knowledge | the match between the knowledge of an expert and a given task |
| 2. Perspective | the expected perspective of the expert, e.g., due to academic background |
| 3. Reliability | the validity, credibility, or soundness of the expert's knowledge based on the expert's competence |
| 4. Up-to-dateness | how recent the expert's knowledge is |
| Accessibility-related factors | |
| 5. Physical proximity | how close or far away the expert is located |
| 6. Availability | the time and effort involved in contacting the expert |
| 7. Approachability | how comfortable the participant feels about approaching the expert |
| 8. Cognitive effort | the cognitive effort involved in understanding and communicating with the expert and processing the obtained information |
| 9. Saves time | how much time the participant saves when contacting this expert |
| Other | |
| 10. Familiarity | whether and how well the participant knows the expert |
| 11. Contacts | the relevance of the expert's contacts |

In a study of trust-related factors in expertise recommendation, Heath, Motta, and Petre (2006) found 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 (2003) showed that knowing about an expert's knowledge, valuing that knowledge, and being able to gain access to an expert's knowledge influence which experts searchers contact for help. Differences between job roles regarding the amount and motivation of expert search, as well as the type of tools used indicate a possible influence of work tasks (Ehrlich & Shami, 2008). The use of social network information is expected to benefit expert search based on domain analysis (Terveen & McDonald, 2005), and users are more likely to select expertise search results that include social network information (Shami, Ehrlich, & Millen, 2008).

Woudstra and Van den Hooff (2008) focused 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 study of information-seeking tasks carried out at participants' workplaces. The study follows a think-aloud protocol, and the importance of individual factors is assessed through counts of how frequently they are mentioned when experts are evaluated. We list the factors identified by Woudstra and Van den Hooff (2008) in Table 1. Quality-related factors (Factors 1–4) appear to be the most important while familiarity (Factor 10) also appears to play a role. We use the factors they identified as the basis for identifying contextual factors in our study.

Further evidence of the usefulness of individual contextual factors (e.g., 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 (Ackerman & McDonald, 2000). 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 (Shami, Yuan, Cosley, Xia, & Gay, 2007). The system uses information on the social network, existing skills, and needed skills of a person, which are provided explicitly by the users. The SmallBlue system mines an organization's electronic communication to provide expert profiling and expertise retrieval (Ehrlich, Lin, & Griffiths-Fisher, 2007). Both textual content of messages and social network information (patterns of communication) are used. The system is evaluated in terms of its usability and utility. Finally, Liebregts and Bogers (2009) detailed the development and systematic evaluation of an expert system. The authors performed system-based, expert-based, and end-user evaluation and showed that combining diverse sources of expertise evidence improves search accuracy and speed.

## Expertise Retrieval

Expertise retrieval aims at developing algorithms that can support the search for expertise using information-retrieval technology. The topic has been addressed at the expert-finding task of the enterprise track, which ran from 2005 to 2008 at the annual Text REtrieval Conference (Bailey, Craswell, de Vries, & Soboroff, 2008; Balog, Soboroff, et al., 2009; Craswell, de Vries, & Soboroff, 2006; Soboroff, de Vries, & Crawell, 2007).

The specific task setup and evaluation of the expert-finding task at TREC have changed over time. The goal was to design a task that reflects real-life expert finding as well as possible while taking into account constraints regarding the available data and resources for relevance assessment. In the first 2 years, the task was based on the W3C collection, consisting of a crawl of the Web site w3c.org, the platform of the


DOI: 10.1002/asi

World Wide Web consortium (W3C, 2005). In 2005, the task was to predict membership in expert groups, given the name of the group as the topic of expertise. Ground truth data for this task was contained in the corpus. In 2006, topics were created and judged by task participants. Topics were similar to ad hoc topics. Systems had to return candidate experts and supporting documents, and expertise was judged by participants based on these supporting documents. In 2007, a new collection was developed based on the intranet pages of the Commonwealth Scientific and Industrial Research Organization (CSIRO; Bailey et al., 2007). The goal of using this collection was to evaluate expertise retrieval and enterprise search in the context of a realistic work task—science communicators of CSIRO developed topics and provided relevance judgments. This setup is closest to the one used in this article. In 2008, the same collection was used, and topics were created and assessed by participants.

Expertise retrieval experiments at TREC have exclusively focused on *expertise finding* (i.e., given a topic, find experts on the topic). Other tasks that have been explored are expert profiling (i.e., given a person, list the areas in which he or she is an expert) and *finding similar experts* (i.e., given a person, return similar experts). In addition to the TREC collections, the UvT collection was introduced by Balog, Bogers, Azzopardi, de Rijke, and van den Bosch (2007). It represents a typical intranet of a large organization. In this setting, a relatively small amount of clean, multilingual data is available for a large number of experts. This collection is used in our study (discussed later).

The task addressed in the current article is finding similar experts, and was first formulated and addressed in Balog and de Rijke (2009): An expert-finding task for which a small number of example experts is given, and the system's task is to return *similar experts*. Balog and de Rijke defined, compared, and evaluated four ways of representing experts: through their collaborations, through the documents with which they are associated, and through the terms with which they are associated (either as a set of discriminative terms or as a vector of term weights). Similarity between experts is computed using the Jaccard coefficient and cosine similarity. Later, we will use these methods and extend them with representations of experts based on self-provided profiles to form our baseline models.

A challenge in content-based expertise retrieval is that systems need to go beyond document retrieval, as they are required to retrieve entities instead of documents. Evidence from documents is used to estimate associations between experts and documents or experts and topics (Balog et al., 2006; Balog, Azzopardi, & de Rijke, 2009). Few algorithms have been proposed that take factors beyond textual document content into account. Extending evidence of expertise beyond the documents directly associated with an expert candidate, Serdyukov, Rode, and Hiemstra (2008) introduced graph-based algorithms that propagate evidence of expertise via several steps of related documents and experts. This approach is similar to the search strategies observed in case studies of expertise seeking: People chain evidence of expertise, finding experts via documents and other experts (Hertzum, 2000). Following a similar intuition, Karimzadehgan, White, and Richardson (2009) leveraged information from graphs representing an organizational hierarchy to enrich information on employees about which little information is known. They showed that people who are close in terms of the organizational hierarchy typically have similar expertise, and that this can be exploited to improve retrieval performance.

Amitay et al. (2008), Balog and de Rijke (2009), Jiang, Han, and Lu (2008), and Serdyukov and Hiemstra (2008) explored the Web as a source of additional expertise evidence. Amitay et al. focused on evidence that can be obtained from Web 2.0 applications such as social bookmarking and blogs. Serdyukov and Hiemstra systematically explored different types of Web evidence considering local, regional, and global evidence, and sources from the general Web, such as specific types of documents, news, blogs, and books. Serdyukov and Hiemstra found that using more sources of evidence typically results in better rankings. Balog and de Rijke found that short text snippets returned by a Web search engine can be used effectively to generate expert profiles.

## Method

In this section, we present the experiments we designed to address our research questions. We first describe the setting in which our experiments were conducted, followed by the data-collection method used to identify contextual factors and to create a test dataset used to perform expertise-retrieval experiments. Next, we describe our expertise-retrieval models: the content-based models that form our baseline, and the principles used to model and integrate contextual factors. Finally, we detail the measures used to evaluate our retrieval models and methods used for parameter estimation.

### *Experimental Setting*

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 6 communication advisors, 1 responsible for the university as a whole and 1 advisor for each of the faculties of Economics and Business Administration, Law, Social and Behavioral Sciences, Humanities, and Theology (cf. Table 2). Typically, the communication advisors receive requests from the media for locating experts on specific topics. Such requests originate from, for example, newspapers and radio shows desiring quick, but informed, reactions to current events, or from 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 is the situation on which we focus in our article: What similar experts should be recommended if the top expert is


DOI: 10.1002/asi

TABLE 2. Overview of communication advisors and candidate experts. Communication Advisors A to E recommend experts from their respective faculties. Advisor F is responsible for communication with experts from any of the faculties. Note that some experts belong to more than one faculty or are not associated with any faculty, so the number of experts in the whole university does not equal the sum of those per faculty.

| Advisor | Faculty | Experts |
|---|---|---|
| A | Economics and Business Administration | 303 |
| B | Humanities | 138 |
| C | Law | 264 |
| D | Social and Behavioral Sciences | 245 |
| E | Theology | 67 |
| F | whole university | 1,168 |

not available, and what factors determine what experts should be recommended?

The document collection we use for our experiments is the existing UvT Expert Collection[2], which was developed for expert finding and expert profiling tasks (Balog et al., 2007). This collection is based on a crawl of an expert-finding system, *WebWijs*, in use at Tilburg University. *WebWijs*[3] is a publicly accessible database of university employees who are involved in research or teaching. This includes, for example, professors, technical and support staff, postdocs and other researchers, and graduate students. Each of the 1,168 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 1,491 knowledge areas, and is encouraged to suggest new knowledge areas that are added upon approval of the *WebWijs* editor. Knowledge areas are organized in a topical hierarchy, and each knowledge area has a separate page devoted to it that shows all experts associated with that area and, if available, a list of related areas. For our experiments, we extended the collection with topics and relevance ratings for the *finding similar experts* task.

A further resource that we use for our experiments is the *media list* available at Tilburg University. This list is compiled annually by the university's Office of Public and External Affairs. This list 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 television appearances receiving 6 points. We considered the media rankings of the 3 years (2005–2007) available to us and collected the average and the total media score for each expert on these lists.

### *Data Collection*

Data collection served two purposes: (a) to gain insights into the factors that play a role when people decide on similar experts in the studied setting and (b) to construct a test set for evaluation of our *similar expert finding* models using standard retrieval-evaluation metrics.

All 6 communication advisors of Tilburg University participated in our study. Data were collected through a printed questionnaire that was filled out by participants in their normal work environment and returned by mail. This setup was chosen because it was deemed to require the least effort for the communication advisors, whose time available for participating in the study was limited. A copy of the questionnaire is provided in the Appendix. Note that the study was conducted in Dutch, the native language of the participants; but here, we provide an English translation.

### *Contextual Factors*

The questionnaire consisted of three parts: (a) background information, (b) relevance assessment, and (c) explicit rating of contextual factors. In the first part, participants were asked for background information about their job function, daily activities, and what information sources they usually consult in their daily activities. They also were asked how often they receive requests for experts, to provide examples of such requests, and to explain how these are typically addressed.

The second part of the questionnaire focused on eliciting relevance judgments for the similar experts task and the factors influencing these decisions. To identify the reasons for participants' relevance decisions, we posed three follow-up questions for each assessed topic: "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 particpants' perceived importance of factors, the third part of the questionnaire asked participants to explicitly rate the overall influence of these factors on their recommendation decisions. We used a 4-point Likert-type scale and the factors identified in Woudstra and Van den Hooff (2008) (discussed earlier, see Table 1). Note that Part 3 of the questionnaire was added only once, at the end of the questionnaire (i.e., after all relevance judgments were made and open questions answered), so that assessors would not be biased by seeing the factors.

### *Retrieval Test Set*

As explained earlier, the second part of the three-part questionnaire used for data collection focused on obtaining relevance judgments for our retrieval test set. In our setting, the test set consists of a set of pairs (target expert, list of similar experts), so "test topics" are experts for whom similar experts need to be found. These test topics were developed as follows. For each communication advisor, we selected the 10 top-ranked employees from their faculty based on the media lists produced by the university's public relations department (discussed earlier). For 1 faculty, the media list contained only 6 employees, and 2 employees were members of two faculties. For the university-wide communication advisor,

[2]http://ilk.uvt.nl/uvt-expert-collection/
[3]http://www.uvt.nl/webwijs/


DOI: 10.1002/asi

the top-10 employees of the entire university were selected.[4] In total, 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 the recommendations and the information sources used or would use to answer this request (cf. Appendix Questions II.1–3). Second (on a separate page of the questionnaire), 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 1 (*least likely to recommend*) to 10 (*most likely to recommend*; cf. Appendix Questions II.4–5). 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, AREAS) described in the next section. Participants were then asked to justify their rating decisions (as described earlier).

We chose to collect the two types of relevance judgments to identify any obvious candidates that the baseline runs were missing. The order of presenting the two questions was chosen to avoid biasing participants by the list of names.

The expert relevance judgments were then constructed in the following way: The ratings supplied by the participants on the 10 experts listed in Appendix Question II.4 were used as the relevance judgments for each test topic. Similar experts who were only mentioned in response to Question II.1, but not in the top-10 list of Question II.4, received the maximum relevance judgment score of 10 (if they were mentioned in both questions, then the rank assigned in Question II.4 was used). Experts who were not rated or not employed by the university anymore were removed. For the 12 duplicate test topics, the ratings by the 2 communication advisors were averaged and rounded to produce a single set of relevance judgments for each topic.

To estimate the reliability of the relevance assessments, we compare assessments on the 12 overlapping topics. Percentage agreement between annotators is 87% if we consider two classes: top-ranked experts (i.e., rated as "10") are considered relevant; all other ratings are considered to be not relevant. In this case, Cohen's $\kappa$ is 0.674, indicating substantial agreement (cf. Landis & Koch, 1977). In addition, in half of the cases, both annotators independently suggested the same expert (i.e., before seeing our suggestion list). This relatively high agreement may indicate that participants can easily identify a small number of similar experts. Agreement at finer granularity (i.e., when considering each rank as one class) is difficult to establish due to low overlap between rankings (Some candidates were not ranked when participants did not feel comfortable rating a candidate), but is generally lower than at the top rank.

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

*Retrieval Models*

*Baselines.* As our baseline approach for finding similar experts, we consider *content-based* similarity only, leaving out any contextual factors. For computing content-based similarity, we represent employees through the content associated with them. We consider two sources: (a) documents associated with these experts and (b) the expertise areas that experts manually selected for their expertise profile in *WebWijs*. These two information sources reflect the two types of sources that are typically available in large organizations and that have previously been used for expert-finding experiments. The first, documents associated with an expert, can be obtained from e-mail or document-versioning systems, or associations can be inferred from people names mentioned in documents. This type of information has been the main focus of experiments in the TREC Enterprise task (Bailey et al., 2007). The second information source is representative of data-warehousing systems where employees have to self-assess their skills (Seid & Kobsa, 2000). Similar information could be mined from publications where authors self-select topical areas from a taxonomy such as the ACM Computing Classification System.[5]

From the two sources of information that we have available, we measure (content-based) similarity of two experts using the following three representations:

$D(e)$: This denotes the set of documents $d$ associated with expert $e$. These documents can be publications, in which case $e$ is an author of $d$, or course descriptions, in which case $e$ is teaching the course described in $d$.

$\vec{t}(d)$: We use $\vec{t}(d)$ to denote 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 defined as the normalized sum of document vectors (for documents authored by $e$): $\vec{t}(e) = \|\sum_{d \in D(e)} \vec{t}(d)\|$.

$K(e)$: This is the set of knowledge areas manually selected by expert $e$ from a finite set of predefined knowledge areas.

Using the representations described previously, we construct the function $sim_T(e, f) \in [0, 1]$ that corresponds to the level of content-based similarity between experts $e$ and $f$ (Table 3). For the set-based representations $[D(e), K(e)]$, we compute the Jaccard coefficient. Similarity between vectors of term frequencies $[\vec{t}(e)]$ is estimated using the cosine distance. The three methods for measuring similarity based on the representations listed earlier are referred to as DOCS, TERMS, and AREAS, respectively. Methods DOCS and TERMS are taken from Balog and de Rijke (2007) while AREAS is motivated by the data made available in *WebWijs*.

As our similarity methods are based on two sources (i.e., documents and knowledge areas), we expect that combinations may lead to improvements over the performance of individual methods. The issue of retrieval run combinations has a long history, and many models have been proposed. We consider one particular choice, Fox and

[5]cf. http://www.acm.org/about/class; similar taxonomies exist for most scientific disciplines.


DOI: 10.1002/asi

TABLE 3. Measuring content-based similarity.

| Method | Data source | Expert representation | Similarity score |
|---|---|---|---|
| DOCS | documents | set: $D(e)$ | $sim_{DOCS}(e, f) = \frac{\lvert D(e) \cap D(f) \rvert}{\lvert D(e) \cup D(f) \rvert}$ |
| TERMS | documents | vector: $\vec{t}(e)$ | $sim_{TERMS}(e, f) = \cos(\vec{t}(e), \vec{t}(f))$ |
| AREAS | knowledge areas | set: $K(e)$ | $sim_{AREAS}(e, f) = \frac{\lvert K(e) \cap K(f) \rvert}{\lvert K(e) \cup K(f) \rvert}$ |

Shaw's (1994) combination combSUM rule, also known as *linear combination*. In this model, overall scores consist of the weighted sum of individual scores. In cases where weights sum to 1, this corresponds to the weighted average of individual scores. Optimal weights are estimated empirically (discussed later).

Given an employee $e$, we compute overall content-based similarity scores $sim_T(e, f)$ (for candidates $f$) as the weighted sum of individual content-based similarity scores $sim_{T_i}$:

$$sim_T(e, f) = \sum_i w_i \cdot sim_{T_i}(e, f), \quad (1)$$

where $w_i$ denotes the weight and $sim_{T_i}$ is the similarity score according to method $T_i$ as defined in Table 3.

*Modeling contextual factors.* On top of the content-based similarity score defined earlier, we model contextual factors that are found to influence human decisions on what similar expert to recommend. Our goal is to optimize the ranking of candidate experts by taking into account additional information from these factors. When we identify which contextual factors play a role in people's decisions on recommending experts, we do not know exactly how these factors influence the decision, and thus how to model them. Therefore, we explore several principled ways of modeling and combining these factors, each of which may reflect ways in which these factors work.

We consider two ways of modeling these factors: (a) *input-dependent* and (b) *input-independent*. In the former case, a candidate $f$'s contextual score depends on features of the person $e$ for whom similar experts are being sought. In the latter case, we compute a contextual score for a candidate expert $f$ independent of the (input) person $e$. In both models, the contextual score of a candidate expert $f$ is combined with the content-based similarity score $sim_T(e, f)$ defined earlier.

**Input-dependent modeling of contextual factors.** Modeling factors in an input-dependent way means that we model a factor as a similarity measure. This option is similar to the content-based baseline models previously discussed in the sense that similarity between experts is estimated using distance in some feature space. The features are designed to reflect the contextual factors that are found to play a role in recommending similar experts (discussed earlier and again in the Results section). For example, one could assume that two employees who are part of the same research group might be recommended as similar experts. A simple model could assign a similarity score of 1 if the two work in the same group, and a score of 0 if they work in different groups.

More formally, given an expert $e$ and a candidate $f$, we determine the input-dependent similarity score $sim_{D_i}(e, f)$ between these experts in terms of the contextual factor $C_i$. For factors with nominal values, we set the similarity to 1 if the values are equal, and to 0 if they differ:

$$sim_{D_i}(e, f) = \begin{cases} 1, & C_i(e) = C_i(f) \\ 0, & otherwise, \end{cases} \quad (2)$$

where $C_i(\cdot)$ denotes the value for the contextual factor $C_i$. In case factor $C_i$ is numeric, we determine the similarity score based on the absolute difference between the normalized values of $C_i$:

$$sim_{Di}(e, f) = 1 - \left| \frac{C_i(e) - C_{i_{min}}}{C_{i_{max}} - C_{i_{min}}} - \frac{C_i(f) - C_{i_{min}}}{C_{i_{max}} - C_{i_{min}}} \right| = 1 - \left| \frac{C_i(e) - C_i(f)}{C_{i_{max}} - C_{i_{min}}} \right|, \quad (3)$$

where $C_{i_{min}}$ and $C_{i_{max}}$ are the minimum and maximum values of factor $C_i$, so that $sim_{D_i}(e, f) \in [0, 1]$.

The final input-dependent contextual similarity score is the linear combination of content-based and contextual similarity scores:

$$sim_D(e, f) = \sum_i w_i \cdot sim_{T_i}(e, f) + \sum_j w_j \cdot sim_{D_j}(e, f), \quad (4)$$

where $w_{i,j}$ denote the weights for individual content-based and contextual similarity methods, as determined through a series of experiments (dicussed later).

**Input-independent modeling of contextual factors.** Here, we model contextual factors independently of the expert $e$ for whom we are seeking similar experts, meaning that we assume that candidates with certain characteristics are more likely to be recommended as a similar expert. For example, a person with a long publication record may be judged to be very reliable, which may increase the overall chance of this person being recommended as an expert. We assume independence of the individual contextual factors $C_i$ and put

$$sim_I(e, f) = \sum_i w_i \cdot sim_{T_i}(e, f) + \sum_j w_j \cdot score_{c_j}(f), \quad (5)$$


DOI: 10.1002/asi

where $w_i$ denotes the weight of the content-based similarity method $i$ and $w_j$ denotes the weight of candidate expert $f$'s score given the value of the contextual factor $C_i$. Note that the assumption of independence is made for simplification, but may not always be realistic.

We model $score_{C_i}(f)$ as the probability of $f$ being recommended as an expert conditioned on $C_i$:

$$score_{C_i}(f) = P[E|C_i(f)] = \frac{P[E \cap C_i(f)]}{P[C_i(f)]}, \qquad (6)$$

where $E$ denotes the event that a candidate $f$ is recommended as an expert, and $P[C_i(f)]$ denotes the probability of observing the value $C_i(f)$ for the specified contextual factor.

The probabilities $P[E \cap C_i(f)]$ and $P[C_i(f)]$ are estimated from the frequencies observed in the data. For nominal contextual factors, obtaining the frequencies is straightforward. For numeric values, we first discretize factors using unsupervised discretization into 10 equal-sized bins and then obtain the counts per bin. Thus, we approximate the probability that a candidate $f$ is recommended as an expert, given an observed value for contextual factor $C_i$ by counting the number of candidates in a specific interval that have been recommended as a similar expert, dividing this by the number of candidates in this interval. To avoid division by zero, we apply the commonly used Laplace estimator (i.e., we initialize all counts to 1).

### *Retrieval Evaluation Metrics*

We use three metrics to evaluate the task of finding similar experts: Expert Coverage (ExCov), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulated Gain (NDCG). 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 (e.g., if the target expert did not select any expertise areas). In recommender systems evaluation, this is typically measured by coverage (Herlocker, Konstan, Terveen, & Riedl, 2004).

MRR is defined as the inverse of the rank of the first retrieved relevant result (in our case, "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. Additionally, our experiments show that human judges achieve high agreement on who they would recommend as the most suitable similar expert (discussed previously). Therefore, we will use MRR as our primary measure of performance.

NDCG is an IR measure that compares ranked lists and 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 (2002), as implemented in trec_eval 8.1.[6]

MRR and NDCG are computed for all experts, including those for which similarity methods resulted in empty lists of recommendations. In other words, "missing names" contribute a value of 0 to all evaluation measures.[7] This allows for a more meaningful comparison between methods, as all scores are calculated based on the same set of test topics.

We evaluate our contextual retrieval models against a baseline consisting of the optimal combination of content-based retrieval models. Significance testing against this baseline is performed using a paired, two-tailed Student's $t$ test. Throughout the article, we use $\Delta$ to indicate runs that significantly outperform the baseline at the 0.05 level.

### *Parameter Estimation and Tuning*

Our goal is to optimize the parameter settings for each model for high MRR (our primary evaluation measure). As is customary when the amount of available training data is small, we optimize and evaluate on the same training set. Note that our goal is not to determine the generalizability of the optimal parameter settings to unseen datasets. Rather, we aim to determine the best possible performance of each model on the available data to determine the upper bound of each model on this data. We then compare these optimized models.

For small numbers of parameters, we obtain optimal settings using parameter sweeps. In cases where the number of parameters to estimate is large (e.g., $>5$), this approach is impractical, as exhaustive search for optimal settings is exponential in the number of parameters. In those cases, we use a simple hill-climbing algorithm to approximate optimal settings in these situations. We randomly initialize the parameters, then vary each parameter between 0 and 1 with increments of 0.1. We select the value for which the target evaluation measure is maximized and then continue with the next parameter. The order in which parameter values are optimized is randomized, and we repeat the optimization process until the settings have converged. Because this algorithm is susceptible to local maxima, we repeat this process 200 times and select the weights that result in the best performance.

## Results

We present our results in three subsections corresponding to our three research questions. First, we analyze the questionnaires we used to identify contextual factors that play a role in the studied setting. We recommend which of these factors should be integrated with content-based retrieval models to address the finding similar experts task, and develop models for these factors. Finally, we present the results of evaluating

[6]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.

[7]This corresponds to running trec_eval with the switch $-$c.


DOI: 10.1002/asi

TABLE 4. Example statements, frequency distribution for statements and participants implicitly mentioning a factor, 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) | Statements ($N = 354$) | Participants ($N = 6$) | *Mdn* 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 |

our models in a retrieval task and analyze the contribution of each factor to the obtained retrieval performance.

### *Identified Contextual Factors*

In this section, we analyze the communication advisors' responses to the questionnaire. We give a short overview of the responses to Part 1 (background information), but then focus on Parts 2 and 3 (identifying contextual factors) of the questionnaire.

The amount of expertise requests typically received and the time spent on answering these requests vary widely between study participants. Half of the participants receive requests one to several times per week. The other half reported receiving requests about once a month. Answering the requests typically takes between 5 and 15 min; however, 1 participant reported that answering complex requests can take up to several hours. Participants use a large variety of sources to keep up with current research within their department. All participants mentioned direct contact with colleagues as an information source. Other sources mentioned are press releases, Web sites of researchers, project descriptions, descriptions of research programs, and the *WebWijs* system.

To identify contextual factors that play a role in expertise recommendations, we analyzed the responses to the open questions of Part 2 of the questionnaire, on why specific recommendation decisions were made. These responses were transcribed and analyzed through content analysis. First, 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 by Woudstra and Van den Hooff (2008); two additional factors, *Organizational structure* and *Media experience* were identified and added to the coding scheme (discussed later). Interannotator agreement was 77.5%, and the chance-corrected agreement Cohen's $\kappa$ was 0.697, indicating substantial agreement (cf. Landis & Koch, 1977). Conflicts were resolved through discussion.

Table 4 gives an overview of the frequency distribution of the resulting factors and the median rating each factor received when participants were asked explicitly to rate these factors. *Topic of knowledge* was mentioned the most often and was mentioned by all participants. 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 of the participants.

Figure 1 allows for a more detailed comparison of factors resulting from coding open responses ("implicit ratings") versus the explicit ratings collected in Part 3 of the questionnaire. There is agreement over all participants 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.

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*.

### *Modeling Factors*

Based on the survey results, we develop recommendations as to which contextual factors should be considered


DOI: 10.1002/asi

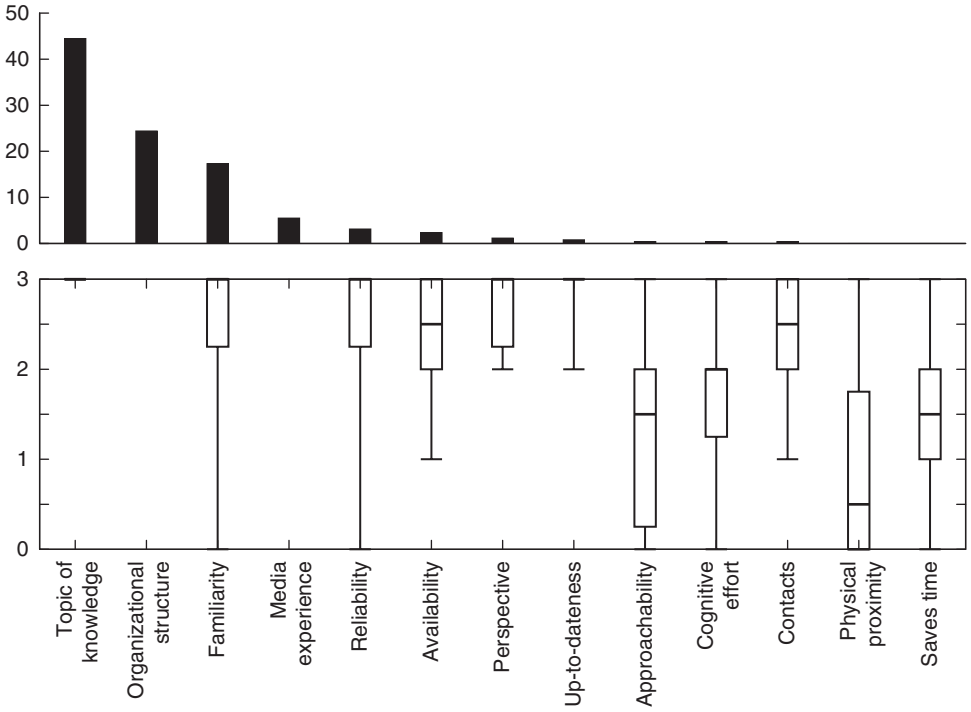


FIG. 1. Frequency of implicit factor mentions (above, in percent) versus explicit ratings (below, on a 4-point Likert-type scale). For explicit ratings median, quartiles, minimum and maximum ratings are indicated. For organizational structure and media experience no explicit ratings are available as these factors were only identified during the analysis of the questionnaires.

for integration in algorithms for finding similar experts in the studied task and environment. *Topic of knowledge*, *Organizational structure*, *Familiarity*, and *Media experience* appear promising, as they received high ratings according to both implicit and explicit measures. Interesting factors are *Up-to-dateness*, *Reliability*, *Availability*, and *Contacts*. Because of the large differences between implicit and explicit ratings 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; factors that will be included in the follow-up experiments are marked with "+" and ones that will not be considered further are marked with "−." For the sake of simplicity, for each contextual factor addressed in this section, we demonstrate the implementation of that factor in one specific way, except for *Reliability* for which we consider both publication record and position.

For the factors that are modeled, we detail their specific implementations $C_i(f)$. Recall that we are considering two types of models: the input-dependent model computes similarity scores between expert candidates and a given candidate based on the value of a contextual factor $C_i(f)$ (cf. Equations 3–5). The input-independent model estimates the probability of a candidate being recommended as a similar expert, again, based on the value of contextual factor $C_i(f)$ (cf. Equations 6 and 7).

\+ *Topic of knowledge* is taken to correspond to the content-based similarity measures presented earlier. This approach represents experts based on (a) the documents associated with them (DOCS and TERMS) and (b) the expertise areas manually selected by the experts themselves (AREAS). We assume that the manually selected labels are the most representative of a person's expertise, as they can be considered as ground truth labels. In addition, participants participating in our user study repeatedly reported using overlap in topic areas as found in the *WebWijs* system as a basis for their (topic-based) decisions on whom to recommend as a similar expert. We include DOCS and TERMS because these have been used to capture expert knowledge in previous work (Balog & de Rijke, 2007).

\+ *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 (Balog et al., 2007).

In our current implementation, we only use the top level of the available organizational hierarchy and consider *organizational structure* as a nominal measure:

$$C_{org}(f) = faculty(f) \tag{7}$$

where *faculty* is a string, for example "FGW" (Faculteit Geesteswetenschappen–Humanities), "FRW" (Faculteit Rechtsgeleerdheid–Legal studies), or "FEB" (Economie en Bedrijfswetenschappen–Economics and Business studies). A staff member may be a member of multiple faculties.


DOI: 10.1002/asi

Thus, in an input-dependent model, this factor expresses whether two staff members are part of the same faculty; in an input-independent model, it expresses how likely someone is to be recommended as an expert, given that they work in a specific faculty.

− *Familiarity* could be implemented in settings where social network information is available, such as patterns of e-mail or other electronic communication (discussed earlier). In our setting, this type of information is currently not available.

+ Information on *media experience* can be obtained from the university's media list (discussed earlier). These media hit counts represent a quantification of media experience and can serve, for instance, as expert priors.

We model the media experience of an expert as the sum of all media scores a candidate has accumulated:

$$C_{media}(f) = \sum_{y} media_y(f), \tag{8}$$

where $media_y(f)$ is the total media appearance score of expert $f$ for year $y$.

+ *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* .

Because both sources of information are readily available, we develop two models for this factor. First, 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 the total number of publications by a candidate $f$:

$$C_{publication}(f) = \sum_{y} pub_y(f), \tag{9}$$

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

A second possibility for assessing an expert's reliability is his or her position within the university or, more generally, the organization. For example, a professor may be more likely to be considered a reliable expert by a communication advisor than may a Ph.D. student. This factor is modeled as nominal. Thus:

$$C_{position}(f) = position(f), \tag{10}$$

where *position* is a string; for example, "Professor," "Lecturer," '"PhD student."

+ *Up-to-dateness* can be modeled by assigning higher weights to more recent documents associated with an expert, such as recent publications. An ideal candidate not only has credible knowledge but this knowledge also is recent. To measure this, we again use the publication records of people, but here, more recent publications receive a higher weight:

$$C_{uptodate}(f) = \sum_{i} \tau^{(y_0 - y_i)} \cdot pub_{yi}(f), \tag{11}$$

where $y_0$ is the current year, and $pub_{y_i}(f)$ is the number of publications of expert $f$ and year $i$. We model the decrease in the influence of older publications using an exponential weight function with base $\theta$. This parameter can be tuned to adjust the rate at which the impact of publications decrease. In our experiments, we set $\theta = 0.7$.

− *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 coauthorship, 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 coworkers' workload (Erickson & Kellogg, 2000).

+ An expert's *contacts* could be modeled by systems that have access to social network information. As we do not have access to this type of data, we model this factor on the basis of coauthored articles and cotaught courses. We assume that the size of the collaboration network is important, so this is what we model.

We consider only the number of coauthors and colecturers; that is, the number of people with which $f$ has coauthored a publication or jointly taught a course. Formally:

$$C_{contacts}(f) = coauth(f), \tag{12}$$

where *coauth(f)* is the number of distinct people with whom $f$ has coauthored a document or colectured a course.

In sum, we model six factors, two nominal ($C_{org}$ and $C_{position}$), and four numeric ($C_{media}$, $C_{publication}$, $C_{uptodate}$, and $C_{contacts}$). Results of integrating these contextual factors with a content-based retrieval system are provided in the next section.

### *Retrieval Performance*

This section contains the results of our models on the retrieval experiment described earlier. We present results for the content-based baseline models, and then for the models integrating contextual factors in both an input-dependent way and in an input-independent way. For each model, we show results for individual factors, and for the optimal combinations of factors.

*Content-based models (Baseline).* Table 5 shows the experimental results for our content-based runs: individual performance of the three individual similarity methods DOCS, TERMS, and AREAS (as listed in Table 3), and weighted combinations of these. We form two combinations: *BL-MRR* is a baseline combination of all content-based methods with weights optimized for MRR, and *BL-NDCG* is a baseline combination optimized for NDCG, as described previously. Later, we compare the contextual models against these two baselines.

Of the three content-based similarity methods, AREAS performs best in terms of both MRR and NDCG. This method achieves an MRR of 0.4, which means that it identifies a correct expert on Rank 2 to 3, on average. The NDCG score is slightly higher than 50% of an optimal ordering. The relatively good performance of AREAS is expected, as this method makes use of the experts' self-provided profiles, which are expected to contain clean data that accurately


DOI: 10.1002/asi

TABLE 5. Factor weights and retrieval results for content-based models. The optimal combinations of all three content-based models form the baselines for subsequent experiments (*BL-MRR* and *BL-NDCG*). Best scores are in bold. Weights are normalized to sum to 1.

| | Content weights | | | | | |
|---|---|---|---|---|---|---|
| Method | DOCS | TERMS | AREAS | %ExCov | MRR | NDCG |
| Optimized for MRR | | | | | | |
| DOCS | 1.0 | – | – | 59.1 | 0.1875 | 0.1718 |
| TERMS | – | 1.0 | – | 100.0 | 0.1740 | 0.3740 |
| AREAS | – | – | 1.0 | 84.1 | 0.4036 | 0.5375 |
| DOCS + TERMS | 0.889 | 0.111 | – | 100.0 | 0.3163 | 0.4522 |
| DOCS + AREAS | 0.889 | – | 0.111 | 88.6 | 0.4615 | 0.5627 |
| TERMS + AREAS | – | 0.667 | 0.333 | 100.0 | 0.5206 | 0.6404 |
| *BL-MRR* | 0.727 | 0.182 | 0.091 | 100.0 | **0.5288** | 0.6619 |
| Optimized for NDCG | | | | | | |
| DOCS + TERMS | 0.9 | 0.1 | – | 100.0 | 0.3161 | 0.4535 |
| DOCS + AREAS | 0.222 | – | 0.778 | 88.6 | 0.4379 | 0.5801 |
| TERMS + AREAS | – | 0.417 | 0.583 | 100.0 | 0.4877 | 0.6899 |
| *BL-NDCG* | 0.091 | 0.273 | 0.636 | 100.0 | 0.5023 | **0.7090** |

captures the topics with which experts are familiar. Problems with this method include data sparseness, for example, when someone did not select any expertise areas. This is reflected in ExCov—for about 15% of the experts, no similar candidates could be identified.

The methods TERMS and DOCS have been used in prior work, and our findings confirm the findings of Balog and de Rijke (2007). TERMS outperforms DOCS by a high margin, according to MRR. We find that DOCS performs well on a small number of topics, but due to sparseness, it does not find any similar experts for a large number of topics. TERMS performs slightly lower on some topics, but due to the high coverage, its average performance is better. Performance in terms of NDCG is similar. In combination with AREAS, both methods improve performance, but improvements are substantially higher with TERMS.

Best scores are achieved with combinations of all three content-based methods. With optimal weights, the content-based methods achieve an MRR of 0.53, corresponding to returning a correct expert at Rank 2, on average. NDCG goes up to 70% of the score that would be achieved by a perfect ranking.

The weights found during the optimization step show different patterns when optimizing for the different performance measures. When optimizing for MRR, the strongest emphasis is put on DOCS; for NDCG, AREAS receives the highest weight.

Figure 2 shows the reciprocal rank for individual topics for the individual scores and for the optimal combination of content-based retrieval methods (BL-MRR). Note that the methods generally perform well: The optimized combination achieves a perfect score on 15 of the 44 test topics. For another 15 topics, the best candidate was returned at Rank 2 or 3. However, there also are 10 topics for which the best candidate was returned at Rank 5 or worse, and for 3 topics, no relevant experts could be retrieved. In some cases, the reason is data sparseness—no topical areas or documents were available for these experts. Additionally, in a small number of cases, knowledge 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. We also see the effects of our simple approach to combination: On seven topics, individual methods achieve a perfect MRR, but in the combined ranking, scores are lower. The most interesting cases are the test topics for which documents and knowledge areas are available, but retrieval scores are still low. In these cases, there must be additional factors that influence human expertise recommendation decisions.

All in all, using content-based methods only, we manage to achieve reasonable retrieval scores, although there is clearly room for improvement. We seek to achieve this improvement by bringing in factors that go beyond the topical relevance of document content.

*Contextual models.*

**Input-dependent modeling of contextual factors.** Table 6 shows the experimental results for combinations of content-based and contextual factors when using input-dependent models. Recall that input-dependent models use contextual factors to calculate a similarity score between an expert and a candidate, analogously to our content-based models. Thus, when we look at individual factors, we rank by whether the candidate has the same position, media experience, reliability score, and so on as the expert.

We combine individual input-dependent models with the three content-based models and separately optimize weights for MRR and NDCG. We also show combinations of all factors, again optimized for both performance measures, and test for significant improvements over the content-based baseline methods (discussed earlier).

We see that adding individual contextual factors improves MRR for *position*, *organizational structure*, *media experience*, and *up-to-dateness*. The largest increase of an individual factor is achieved with *organizational structure*,


DOI: 10.1002/asi

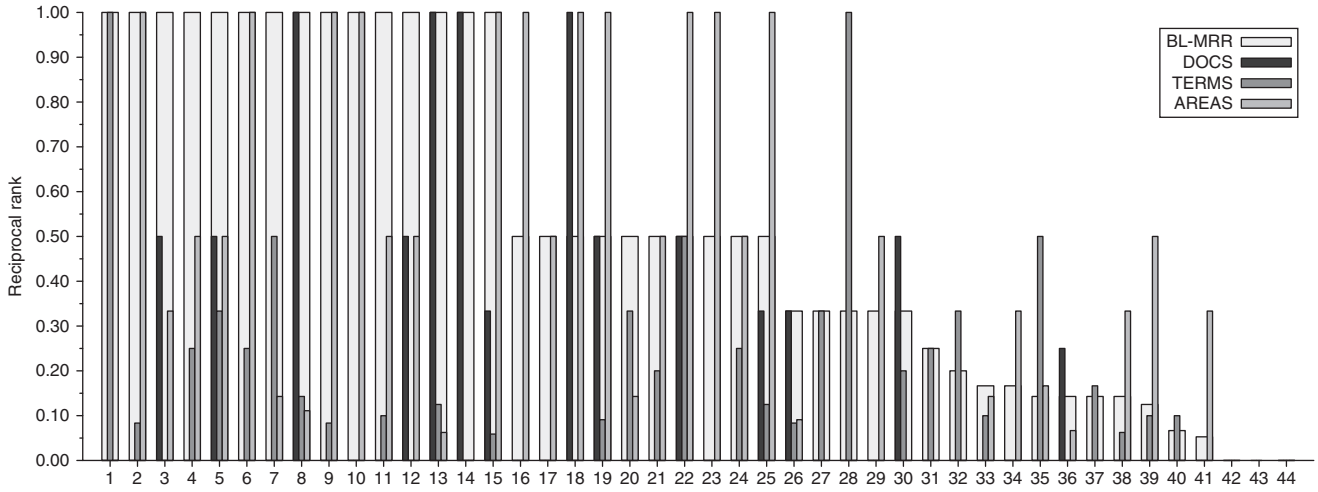


FIG. 2. Per-topic reciprocal rank for the individual (DOCS, TERMS, AREAS) and combined (BL-MRR) content-based methods, sorted by reciprocal rank of BL-MRR. High scores are achieved on many topics, but there is room for improvement on topics where none of the content-based runs achieves high scores.

TABLE 6. Factor weights and retrieval results for *input-dependent* models of contextual factors. Coverage (ExCov) is 100% for all combinations. In the Method column POSITION + BL refers to a combination of the content-based factors and the feature *position* as detailed in the "Modeling Factors" section; similarly for ORG (*organizational structure*), MEDIA (*media experience*), PUB (*publication record*), UPTODATE (*up-to-dateness*), and CONTACTS (*contacts*).

| | Contextual factor weights | | | | | | Content weights | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | position | organizational structure | media experience | publication record | up-to-dateness | contacts | DOCS | TERMS | AREAS | MRR | NDCG |
| Optimized for MRR | | | | | | | | | | | |
| BL-MRR | – | – | – | – | – | – | 0.727 | 0.182 | 0.091 | 0.5288 | 0.6619 |
| POSITION + BL | 0.050 | – | – | – | – | – | 0.450 | 0.050 | 0.450 | 0.5440 | 0.6307 |
| ORG + BL | – | 0.083 | – | – | – | – | 0.667 | 0.167 | 0.083 | 0.5514 | 0.6450 |
| MEDIA + BL | – | – | 0.094 | – | – | – | 0.312 | 0.281 | 0.312 | 0.5436 | 0.7162 |
| PUB + BL | – | – | – | – | – | – | 0.727 | 0.182 | 0.091 | 0.5288 | 0.6619 |
| UPTODATE + BL | – | – | – | – | 0.091 | – | 0.636 | 0.182 | 0.091 | 0.5359 | 0.6585 |
| CONTACTS + BL | – | – | – | – | – | – | 0.727 | 0.182 | 0.091 | 0.5288 | 0.6619 |
| ALL FACTORS | 0.043 | 0.043 | 0.043 | – | 0.087 | – | – | 0.435 | 0.348 | **0.5569** | 0.6287 |
| Optimized for NDCG | | | | | | | | | | | |
| BL-NDCG | – | – | – | – | – | – | 0.091 | 0.273 | 0.636 | 0.5023 | 0.7090 |
| POSITION + BL | – | – | – | – | – | – | 0.091 | 0.273 | 0.636 | 0.5023 | 0.7090 |
| ORG + BL | – | – | – | – | – | – | 0.091 | 0.273 | 0.636 | 0.5023 | 0.7090 |
| MEDIA + BL | – | – | 0.056 | – | – | – | 0.167 | 0.222 | 0.556 | 0.5107 | **0.7360**$^{\Delta}$ |
| PUB + BL | – | – | – | 0.048 | – | – | 0.143 | 0.333 | 0.476 | 0.4956 | 0.7242 |
| UPTODATE + BL | – | – | – | – | 0.056 | – | 0.111 | 0.278 | 0.556 | 0.4865 | 0.7274 |
| CONTACTS + BL | – | – | – | – | – | 0.071 | 0.071 | 0.357 | 0.500 | 0.4915 | 0.7149 |
| ALL FACTORS | – | – | 0.056 | – | – | – | 0.167 | 0.222 | 0.556 | 0.5107 | **0.7360**$^{\Delta}$ |

where an MRR of 0.5514 is achieved. When combining all factors, we achieve an MRR of 0.5569. In this combination, the weights of the contextual factors constitutes 21.6% of the total weight.

When optimizing for NDCG, individual factors that improve performance over content-based methods are *media experience*, *publication record*, *up-to-dateness*, and *contacts*. Combining all factors does not improve over the combination of *media experience* with content-based methods. This combination achieves an NDCG score of 0.736, which is significantly better than the content-based baseline. In this combination, *media experience* receives 5.6% of the total weight.

**Input-independent modeling of contextual factors.** Table 7 shows the performance of input-independent models. As for input-dependent models in the previous section, we first combine individual contextual factors with the content-based models, and then generate combinations of all factors, optimizing both for MRR and NDCG, respectively.

When optimizing for MRR, all individual factors except *organizational structure* improve performance over the content-based baseline model. The largest improvement of an individual factor is obtained with *media experience*, but using all factors further improves on that combination.

When optimizing for NDCG, we find that all individual contextual factors significantly improve over the


DOI: 10.1002/asi

TABLE 7. Factor weights and retrieval results for *input-independent* models of contextual factors. Coverage (ExCov) is 100% for all combinations.

| | Contextual factor weights | | | | | | Content weights | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | position | organizational structure | media experience | publication record | up-to-dateness | contacts | DOCS | TERMS | AREAS | MRR | NDCG |
| Optimized for MRR | | | | | | | | | | | |
| BL-MRR | – | – | – | – | – | – | 0.727 | 0.182 | 0.091 | 0.5288 | 0.6619 |
| POSITION+BL | 0.250 | – | – | – | – | – | 0.050 | 0.250 | 0.450 | 0.5612 | 0.7029 |
| ORG + BL | – | – | – | – | – | – | 0.727 | 0.182 | 0.091 | 0.5288 | 0.6619 |
| MEDIA + BL | – | – | 0.121 | – | – | – | 0.303 | 0.273 | 0.303 | 0.5735 | 0.6725 |
| PUB + BL | – | – | – | 0.067 | – | – | – | 0.600 | 0.333 | 0.5536 | 0.6952 |
| UPTODATE + BL | – | – | – | – | 0.067 | – | – | 0.600 | 0.333 | 0.5372 | 0.6890 |
| CONTACTS + BL | – | – | – | – | – | 0.187 | 0.062 | 0.312 | 0.437 | 0.5642 | 0.6928 |
| ALL FACTORS | 0.107 | 0.214 | 0.036 | – | – | 0.107 | 0.036 | 0.25 | 0.25 | **0.6070** | 0.6452 |
| Optimized for NDCG | | | | | | | | | | | |
| BL-NDCG | – | – | – | – | – | – | 0.091 | 0.273 | 0.636 | 0.5023 | 0.7090 |
| POSITION + BL | 0.050 | – | – | – | – | – | 0.100 | 0.350 | 0.500 | 0.5110 | **0.7452**$^{\Delta}$ |
| ORG + BL | – | 0.056 | – | – | – | – | 0.222 | 0.167 | 0.556 | 0.4974 | 0.7369$^{\Delta}$ |
| MEDIA + BL | – | – | 0.050 | – | – | – | 0.050 | 0.400 | 0.500 | 0.5277 | 0.7353$^{\Delta}$ |
| PUB + BL | – | – | – | 0.053 | – | – | 0.105 | 0.316 | 0.526 | 0.5129 | 0.7334$^{\Delta}$ |
| UPTODATE + BL | – | – | – | – | 0.050 | – | 0.050 | 0.400 | 0.500 | 0.4903 | 0.7289$^{\Delta}$ |
| CONTACTS + BL | – | – | – | – | – | 0.053 | 0.053 | 0.368 | 0.526 | 0.5317 | 0.7369$^{\Delta}$ |
| ALL FACTORS | 0.050 | – | – | – | – | – | 0.100 | 0.350 | 0.500 | 0.5110 | **0.7452**$^{\Delta}$ |

content-based baseline. Like for input-dependent models, the combination of content-based models plus *position* performs best.

The improvements when using input-independent contextual factors indicate that there are certain characteristics that are related to a candidate being more likely to be recommended as an expert. For example, we found that for *position* "Professors" are the most likely to be recommended as an expert while "PhD students" are the least likely to be recommended.

Between the two approaches to modeling contextual factors, we find that the input-independent model performs better for most factors. When optimizing for MRR, this is the case for all factors except *organizational structure*. In addition, in this case, the weights given to input-independent models are substantially higher than for input-dependent models. The combination of all content-based and input-independent contextual factors assigns almost half of the weight mass (46.4%) to the contextual factors.

When optimizing for NDCG, performance is again better for input-independent models, except for *media experience*, where we see slightly better performance with the input-dependent model. The weights when optimizing for NDCG are similar across the two types of models.

**Combining input-dependent and input-independent models.** Finally, we present the retrieval results when combining input-dependent and input-independent models. Intuitively, we expect such a combination to perform best, selecting, for example, a candidate with a high media score who is in the same department as the expert for which recommendations are sought. Table 8 shows the resulting scores.

The best MRR score is achieved by combining the content-based factors with input-dependent *organizational structure*, and input-independent *position*, *organizational structure*, *media experience*, and *contacts*. This combination achieves an MRR of 0.6248, which is significantly better than the baseline.

The best NDCG score is achieved by the combination of input-independent *position* with the content-based models. Like the model optimized for MRR, this combination performs significantly better than does the content-based baseline.

After looking at overall performance of the retrieval models, we now zoom in to per-topic performance. Figure 3 shows per-topic reciprocal rank scores of the baseline model and the combination of all factors as specified in Table 8, and the per-topic difference in reciprocal rank between the two models.

We see that the combination of all factors lost performance on 8 topics, but improved performance on 16 topics. While the baseline model achieved a perfect score on 15 topics, the contextual model achieves this score on 23 topics. These performance differences are mostly due to re-ranking. For example, in Topic 31, the first relevant candidate is moved from Rank 4 to Rank 1. In this topic, the new model also adds two additional relevant experts to the result list, but these are added at lower ranks and do not influence the performance score. In Topic 39, both models retrieved all relevant results, but the new model moved the first relevant result from Place 8 to the top of the result list.

### *Regression Analysis*

In the previous section, we found that our models of contextual factors can lead to significant improvements in retrieval performance. To get a more detailed picture of the relation between the individual factors and relevance assessments, we conduct a regression analysis. This analysis tells


DOI: 10.1002/asi

TABLE 8. Factor weights and retrieval results for combinations of *input-dependent* and *input-independent* models of contextual factors. Coverage (ExCov) is 100% for all combinations.

| | Contextual factor weights | | | | | | Content weights | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | position | organizational structure | media experience | publication record | up-to-dateness | contacts | DOCS | TERMS | AREAS | MRR | NDCG |
| BL-MRR | | | | | | | | | | | |
| input-dependent | – | – | – | – | – | – | | | | | |
| input-independent | – | – | – | – | – | – | 0.727 | 0.182 | 0.091 | 0.5288 | 0.6619 |
| ALL FACTORS | | | | | | | | | | | |
| input-dependent | – | 0.029 | – | – | – | – | | | | | |
| input-independent | 0.171 | 0.089 | 0.029 | – | – | 0.114 | 0.029 | 0.286 | 0.257 | **0.6248**$^{\Delta}$ | 0.6505 |
| BL-NDCG | | | | | | | | | | | |
| input-dependent | – | – | – | – | – | – | | | | | |
| input-independent | – | – | – | – | – | – | 0.091 | 0.273 | 0.636 | 0.5023 | 0.7090 |
| ALL FACTORS | | | | | | | | | | | |
| input-dependent | – | – | – | – | – | – | | | | | |
| input-independent | 0.050 | – | – | – | – | – | 0.100 | 0.350 | 0.500 | 0.5110 | **0.7452**$^{\Delta}$ |

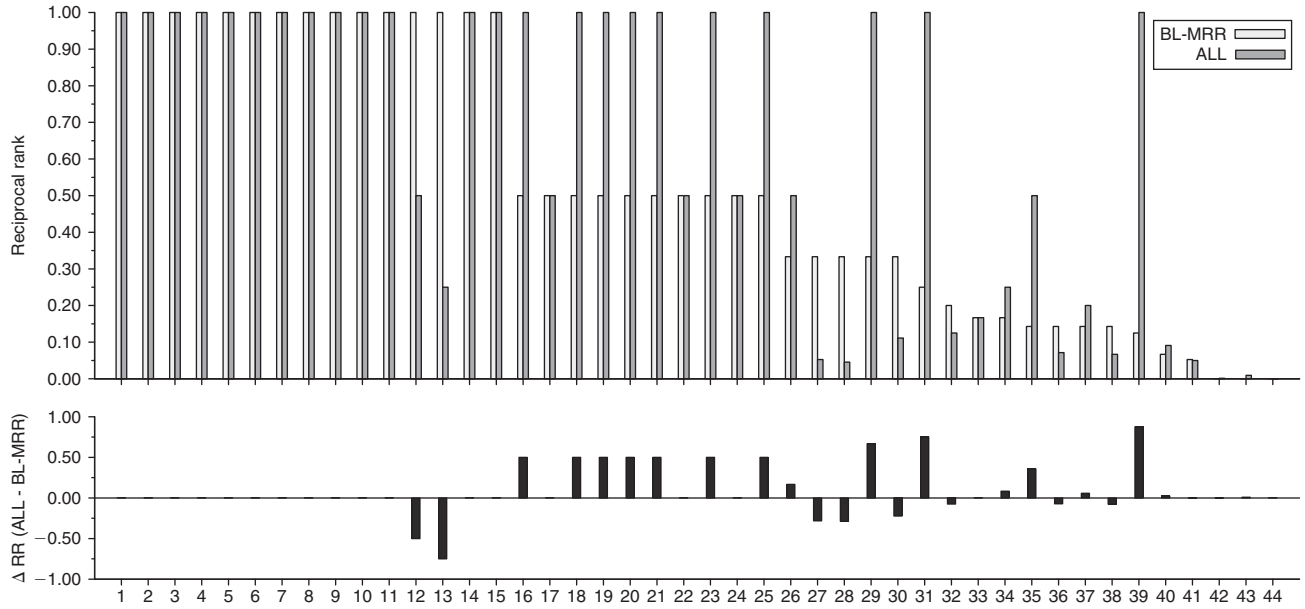


FIG. 3. Per-topic reciprocal rank (top) and difference in reciprocal rank (bottom) for the content-based baseline run (BL-MRR) and the optimized run using all input-dependent and input-independent factors.

us what the contribution of that factor is toward explaining variability in assessments.

We use logistic regression, where we model the problem of predicting the probability of a candidate being judged relevant (i.e., we use the binary assessments, same as for MRR) given an input vector that consists of all factors explored in this article. The results can be seen in Table 9.

From the regression analysis, we first look at the *p*-values. The factors that obtain a small *p*-value have a significant impact on the modeled result (i.e., relevance decisions). Including the factor improves the model's predictions, and the significance of the performance improvement is indicated.

Significant performance improvements are observed for the content-based factors TERMS and AREAS, for the input-dependent model of *organizational structure*, and for the input-independent models of *position*, *organizational structure*, *media experience*, and *contacts*. The *p* level is highest for the factors that were the most frequently mentioned in the user study (*topic*, *organizational structure*, *media experience*)—these are significant at the level $p < 0.001$.

We also list the beta coefficients, which give an indication of the relative weight of each factor, normalized for the variance of the factor and the overall dataset. We see that the highest beta coefficient is assigned to input-dependent *organizational structure*, followed by *position* and input-independent *organizational structure*.

In this section, we identified a number of contextual factors that appear to play a role in finding similar experts in the context of media communications at a university. We developed models for the most interesting factors and integrated them with existing, content-based retrieval models. Our results show that such contextual factors play a role in


DOI: 10.1002/asi

TABLE 9. Output of analyzing factors using regression analysis. We use a logistic model, trying to predict binary assessments from all factors. We report standardized (beta) coefficients $\beta$- and $p$-values for each factor.

| Factor | $\beta$ | $p$ |
|---|---|---|
| Content-based models | | |
| DOCS | 0.4500 | 0.3291 |
| TERMS | 4.9980 | 0.0000*** |
| AREAS | 7.2828 | 0.0000*** |
| Input-dependent models | | |
| *position* | 5.5743 | 0.1174 |
| *organizational structure* | 12.4106 | 0.0001*** |
| *media experience* | −3.9030 | 0.2656 |
| *publication record* | −6.7375 | 0.1695 |
| *up-to-dateness* | 8.7424 | 0.0856 |
| *contacts* | 8.1743 | 0.1138 |
| Input-independent models | | |
| *position* | 11.6155 | 0.0095** |
| *organizational structure* | 10.5065 | 0.0224* |
| *media experience* | 6.0796 | 0.0001*** |
| *publication record* | 1.0874 | 0.7261 |
| *up-to-dateness* | 0.9338 | 0.7292 |
| *contacts* | 7.5787 | 0.0268* |

Contributions identified as significant are marked using ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

selecting experts in the studied situation, and that applying these factors to retrieval can result in significant improvements in retrieval performance. In addition, we conducted a regression analysis to gain a better understanding of which factors contribute to the observed performance improvements. Further implications are discussed in the next section.

## Discussion

In this section, we revisit our original three research questions. First, we address the question of which contextual factors play a role in the studied setting. Next, we discuss the models of contextual factors that were implemented in our study, and finally, the retrieval performance of these models.

### *Identifying Relevant Contextual Factors*

Our first goal was to identify contextual factors that play a role in the task of finding similar experts in response to media requests for expertise. Identified factors were based on a study where different sets of questions were aimed at eliciting both implicit and explicit feedback on what factors study participants deemed important. After modeling the factors and evaluating their performance in retrieval experiments, we can revisit the factors and compare their relative impact with the results obtained from the initial study.

Based on the questionnaire results, we selected those factors for further exploration that were ranked high according to explicit rating, or that were frequently mentioned implicitly, in reasons for explaining recommendation decisions. On some factors, these two measures agree (e.g., *topic of knowledge*, *familiarity*, and *reliability* rank high according to both measures; *saves time* and *physical proximity* are ranked low according to both measures). On other factors, there is disagreement (e.g., on *up-to-dateness* and *perspective*).

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 our study and the study of Woudstra and Van den Hooff (2008), from which we took our coding scheme. In our case, the task was to recommend an expert to a media representative; in the study of Woudstra and Van den Hooff, the experts were assumed to be sought by the participants themselves. It appears that participants take these task characteristics into account. Similarly, organizational structure may not have played a role in the tasks considered in Woudstra and Van den Hooff. 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.

Apart from the two new factors, the frequency distribution of implicit factor mentions is similar to those obtained by Woudstra and Van den Hooff (2008). In both studies, *topic of knowledge* is the most frequently mentioned factor (44.5% in our study; 50–52% in Woudstra and Van den Hooff). *Familiarity* is frequently mentioned (17.3 vs. 8–18%, respectively). Factors relating to accessibility (*physical proximity*, *availability*, *approachability*, *cognitive effort*, *saves time*) are consistently mentioned with very low frequency. Two differences are that our study found fewer instances mentioning *reliability* (3.1 vs. 9%, respectively) and *perspective* (1.2 vs. 9–15%, respectively). Differences can be attributed to differences in task and format of the study.

The importance of the factors mentioned in our study may vary between faculties and between communication advisors. For example, 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 it was 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 plays a role. A more detailed analysis of the differences between faculties is beyond the scope of this article, but it would be interesting to further explore this aspect in future work.

Besides the results of the user study, the relative importance of factors can be assessed based on the results of our retrieval experiments. Of the content-based similarity measures, AREAS always performs best. This is expected because it corresponds to the overlap in knowledge areas that experts themselves have selected. Problems in using this measure for finding similar experts include data sparseness—some experts may have selected topics where they are the sole expert. In addition, a topic area may be too broad


DOI: 10.1002/asi

(e.g., "History"), in which case too many candidates would be retrieved. Similar problems appear to play a role in DOCS and TERMS. For candidates with few publications in the database, there may be little overlap with other expert candidates. For candidates with many publications, overlap may be large, which results in a large number of candidates to be retrieved; that is, the result set will be too broad.

Looking at models integrating individual contextual factors, performance improved with almost all factors. The highest MRR for input-dependent models was achieved by *organizational structure*, followed by *position* and *media experience*. Performance for input-independent models is higher overall, and the highest MRR is achieved by *contacts*, followed by *media experience* and *position*. This higher performance indicates that the factors play a role, but it does not say anything about the relative impact of the factors.

In linear combinations of several factors, we could assume that the relative weights of the factors give an indication of their relative impact on the retrieval results. However, as described earlier, our combinations may suffer from local optima, and weights vary widely, for example, when optimizing for different measures. Therefore, weights can give some indication of what factors played a role, but the magnitude may be misleading.

A regression analysis was conducted to gain further insights into how individual factors contribute to improvements in retrieval performance. This analysis identified several content-based and contextual factors that make a significant contribution in explaining variability in users' assessments of expert similarity. Of the content-based baseline methods, AREAS and TERMS were found to have a significant impact. For the contextual factors, the input-dependent model of *organizational structure* and the input-independent model of *media experience* most strongly contributed to correct predictions of relevance assessments, followed by the input-independent models of *position*, *organizational structure*, and *contacts*. This model is very similar to the factors included in the combination of all factors in Table 8, suggesting that our combined ranking was indeed reasonable.

Most modeled factors were found to have a significant impact, except for *up-to-dateness* and *publication record*. For the factors that did not have a significant influence and that were not included in the retrieval model, there can be several explanations for their limited contribution. First, it is possible that these factors do not really have an influence on the relevance judgments. For *up-to-dateness*, the explicit ratings from the user study were very high, but the factor was mentioned rarely. Study participants may have overestimated the influence of this factor when asked for explicit ratings. Another possibility is that our models do not capture the important aspects of these factors as intended. A future task would be to develop alternative models of *up-to-dateness*. This factor was particularly difficult to model, as publication records capture previous work but not current work. Someone may have just started working in an area and may have no current publications on a topic but still have a lot of knowledge about the area. *Publication record* was one of two possible ways in which we modeled reliability. We conclude that this factor is better captured through *position*.

To summarize, we find multiple pieces of evidence that indicate that besides the *topic of knowledge* as modeled by the content-based baseline, *organizational structure*, *media experience*, and *position* play a role in finding similar experts in the studied setting. All three factors are rated high according to the user study, their models achieve high performance improvements when added to the baseline, and regression analysis shows a significant impact on relevance ratings. We have similar evidence for *contacts*, but performance improvements using this factor are not as high, and regression analysis results in a much higher $p$ value than that for other factors. For *up-to-dateness* and *publication record*, we do not find a significant impact.

### *Modeling Factors*

Our second research question was how to model contextual factors and integrate them with existing retrieval models. In the previous sections, we have detailed two principled models of contextual factors. We explored modeling factors as input-dependent, similar to content-based similarity methods, and as input-independent, similar to a prior probability. The intuition between the first model is that candidates with similar characteristics to the given target expert would be likely to be recommended. The intuition behind the second model is that there may be certain characteristics that make a candidate more likely to be recommended, independent of the target expert.

We found that both types of models improved upon the baseline using content-based factors only. Overall, input-independent models led to better performance, except for the input-dependent model of *organizational structure*. Thus, for the studied setting, it is important that a candidate expert is part of the same department as the topic expert; but in addition to that, there are attributes that are common to frequently recommended experts, such as having prior media experience or being a professor. Best performance was achieved with a run that combined both types of models. These results show that both types of models are useful and that it is not enough to identify a factor but that it also needs to be modeled appropriately. We have explored two types of models, but others may be useful and should be explored in the future.

### *Retrieval Performance*

The third question was whether integrating contextual factors with content-based retrieval methods would improve retrieval performance. Our results show that our models that include contextual factors indeed achieve significant improvements over the content-based baseline methods. Three limitations regarding our retrieval experiments have to be addressed: the method for combining and tuning models, the experimental setup, and the choice of evaluation measure.

The method for combining factors that was used in this work was relatively simple, following accepted practice.


DOI: 10.1002/asi

We formed weighted linear combinations and tuned weights using parameter sweeps where feasible and hill-climbing otherwise. It is very likely that this combination does not optimally fit the distributions of the different factors, and our current approach suffers from local optima. We expect that much better performance can be achieved with more elaborate methods such as current approaches for learning to rank. For the purpose of this article, we are satisfied with showing that retrieval performance can be improved when integrating contextual factors, and we were not focusing on tuning our models to achieve the best possible performance. In the future, it would be interesting to explore other methods for combination.

One of the reasons more complex combinations were not feasible for the current work is the size of the dataset available for our retrieval experiments. While the size of our dataset matches that of a typical TREC setting, it is too small to be able to apply machine learning in an effective manner. Therefore, we limit ourselves to weighted linear combinations, and the weights are optimized on the same dataset.

Finally, note that the optimal weights that achieve best scores on MRR and NDCG vary widely in most combinations. The combinations that work well for MRR do not work well for the other measure. Therefore, the choice of evaluation measure has an influence on the results. It remains an open question as to which measure we should optimize for. In this work, we chose MRR as the main measure, mainly because (a) we found it to be assessed more reliably by assessors and because (b) we think that for this task returning a few good candidates in the top ranks is more important than having a long list of correctly ranked candidates. Other measures may be more appropriate in other tasks, and it would be interesting to look at how the measures relate to each other and to end-user preference.

### *Conclusion*

In this article, we started from the observation that contextual factors appear to play a role in expertise seeking. We explored the role of contextual factors in the task of finding similar experts. First, we identified contextual factors that play a role in the task of finding similar experts in the public relations department of a university. The identified factors were modeled in two principled ways and implemented using available data. We integrated the resulting models with existing, content-based models and evaluated them to assess retrieval performance.

We found that while *topic of knowledge* appears to be the most important factor in the studied setting, contextual factors play a role as well (e.g., *organizational structure*, *position*, *media experience*, and *contacts*). Implementing contextual factors and integrating them with content-based retrieval algorithms resulted in improved retrieval performance. Of the two principled models of contextual factors that we explored, input-independent models performed better than did input-dependent models, but a combination of both types of models performed best overall in terms of MRR, significantly outperforming a competitive baseline. Other models are possible, and more elaborate combinations of different models may be a further promising direction for future work. We plan to explore other ways of integrating contextual factors with content-based retrieval models.

The individual contextual factors that appear to have the most impact are *media experience*, *organizational structure*, and *position*. This finding suggests that there may be a strong task-specific component to the contextual factors that play a role in finding similar experts, and possibly in other retrieval tasks as well. In future work, it would be interesting to perform similar studies of contextual factors in information-seeking tasks in other settings. Based on findings from several such studies, it may be possible to develop more general models of how tasks relate to other factors, and how these relations influence people's relevance decisions.

Overall, our results indicate that identifying contextual factors and integrating them with content-based expertise retrieval models is indeed a promising research direction. 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.

We end on two more general notes. First, this article was concerned with the issue of locating information intermediaries—experts. The more general issue of (access to) high-quality information is increasingly receiving attention from the information science and information retrieval communities. Applications include selecting experts for peer review (Karimzadehgan, Zhai, & Belford, 2008), forming multidisciplinary teams (Rodrigues, Oliveira, & de Souza, 2005), and many others, and we expect that beyond the textual content of documents associated with candidates, contextual factors play a role in these settings as well. In the setting of user-generated content, there is a growing body of computational work on credibility, for example, in blogs (Weerkamp & de Rijke, 2008) and in podcasts (Tsagkias, Larson, Weerkamp, & de Rijke, 2008). What is lacking so far in this line of work is a solid grounding of the credibility features in actual access tasks and scenarios of the type illustrated in this article—we intend to adopt and apply the article's methods to those areas.

Finally, in information-seeking research, models of how contextual factors play a role have been developed, and it has been shown that information-seeking behavior changes with, for example, specifics of the task (Byström & Järvelin, 1995; Kim, 2009) and the user's problem stage (Vakkari, 2001). From an information-retrieval perspective, these contextual factors are difficult to model, and researchers typically design experiments where they abstract from context to make results generalizable. In this article we have argued that to arrive at generalizable results, we need to model context and develop models of how contextual factors influence expertise seeking. We have shown that the factors can be modeled, that it is possible to integrate them with retrieval models, and that the resulting models can improve retrieval performance.


DOI: 10.1002/asi

## Acknowledgments

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. We thank Lilian Woudstra for her extensive feedback and Joost Kircz for discussions and feedback.

This research was supported by the Netherlands Organization for Scientific Research (NWO) under Project Nos. 017.001.190, 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802; by the Dutch and Flemish Governments (http://www.stevin-tst.org) under projects DAESO (STE-05–24) and DuOMAn (STE-09–12); and by the IOP-MMI program of SenterNovem/The Dutch Ministry of Economic Affairs, as part of the À Propos project.

## References

Ackerman, M.S., & McDonald, D.W. (2000). Collaborative support for informal information in collective memory systems. Information Systems Frontiers, 2(3–4), 333–347.

Amitay, E., Carmel, D., Golbandi, N., Har'El, N., Ofek-Koifman, S., & Yogev, S. (2008). Finding people and documents, using web 2.0 data. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008) Workshop on Future Challenges in Expertise Retrieval (fCHER) (pp. 1–5) New York: ACM Press.

Bailey, P., Craswell, N., de Vries, A.P., & Soboroff, I. (2008). Overview of the TREC 2007 Enterprise Track. In Proceedings of the 16th Text Retrieval Conference. Retrieved January 11, 2010, from http://trec.nist.gov/pubs/trec16/papers/ENT.OVERVIEW16.pdf

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

Balog, K. (2008). People search in the enterprise. Unpublished doctoral thesis, University of Amsterdam.

Balog, K., Azzopardi, L., & de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006) (pp. 43–50). New York: ACM Press.

Balog, K., Azzopardi, L., & de Rijke, M. (2009). A language modeling framework for expert finding. Information Processing & Management, 45(1), 1–19. DOI:10.1016/j.ipm.2008.06.003

Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., & van den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007) (pp. 551–558). New York: ACM Press.

Balog, K., & de Rijke, M. (2007). Finding similar experts. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007) (pp. 821–822). New York: ACM Press.

Balog, K., & de Rijke, M. (2009). Combining candidate and document models for expert search. In Proceedings of the 17th Text Retrieval Conference. Retrieved January 11, 2010, from http://trec.nist.gov/pubs/trec17/papers/uamsterdam-derijke.ent.rev.pdf

Balog, K., Soboroff, I., Thomas, P., Craswell, N., de Vries, A.P., & Bailey, P. (2009). Overview of the TREC 2008 enterprise track. In Proceedings of the 17th Text Retrieval Conference. Retrieved January 11, 2010, from http://trec.nist.gov/pubs/trec17/papers/ENTERPRISE.OVERVIEW.pdf

Borgatti, S.P., & Cross, R. (2003). A relational view of information seeking and learning in social networks. Management Science, 49(4), 432–445.

Byström, K., & Järvelin, K. (1995). Task complexity affects information seeking and use. Information Processing & Management, 31(2), 191–213.

Constant, D., Sproull, L., & Kiesler, S. (1996). The kindness of strangers: The usefulness of electronic weak ties for technical advice. Organization Science, 7(2), 119–135.

Cool, C., & Spink, A. (2002). Issues of context in information retrieval (IR): An introduction to the special issue. Information Processing & Management, 38(5), 605–611.

Craswell, N., de Vries, A.P., & Soboroff, I. (2006). Overview of the TREC-2005 Enterprise Track. In Proceedings of the 14th Text Retrieval Conference. Retrieved January 11, 2010, from http://trec.nist.gov/pubs/trec14/papers/ENTERPRISE.OVERVIEW.pdf

Davenport, T.H., & Prusak, L. (1998). Working knowledge: How organizations manage what they know. Boston: Harvard Business School Press.

ECSCW'99 Workshop. (1999). Beyond knowledge management: Managing expertise. Retrieved January 11, 2010, from http://www.iai.uni-bonn.de/~prosec/ECSCW-XMWS/

Ehrlich, K., Lin, C.-Y., & Griffiths-Fisher, V. (2007). Searching for experts in the enterprise: Combining text and social network analysis. In Proceedings of the ACM Conference on Supporting Group Work (GROUP 2007) (pp. 117–126). New York: ACM Press.

Ehrlich, K., & Shami, S.N. (2008). Searching for expertise. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008) (pp. 1093–1096). New York: ACM Press.

Erickson, T., & Kellogg, W. (2000). Social translucence: An approach to designing systems that support social processes. ACM Transactions on Computer–Human Interaction, 7(1), 59–83.

Fox, E.A., & Shaw, J.A. (1994). Combination of multiple searches. In D.K. Harman (Ed.), Proceedings of the Second Text Retrieval Conference (TREC 2). Retrieved January 11, 2010, from http://trec.nist.gov/pubs/trec2/papers/txt/23.txt

Heath, T., Motta, E., & Petre, M. (2006). Person to person trust factors in word of mouth recommendation. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2006) Workshop on Reinventing Trust, Collaboration, and Compliance in Social Systems (Reinvent 06). Retrieved January 11, 2010, from http://eprints.aktors.org/446/01/heath-motta-petre-reinvent2006-person-to-person-trust-factors.pdf

Herlocker, J.L., Konstan, J.A., Terveen, L.G., & Riedl, J.T. (2004). Evaluating collaborative filtering recommender systems. ACM Transactions on Information Systems, 22(1), 5–53.

Hertzum, M. (2000). People as carriers of experience and sources of commitment: Information seeking in a software design project. New Review of Information Behavior Research, 1(January), 135–149.

Hofmann, K., Balog, K., Bogers, T., & de Rijke, M. (2008). Integrating contextual factors into topic-centric retrieval models for finding similar experts. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008) Workshop on Future Challenges in Expertise Retrieval (fCHER) (pp. 29–36). New York: ACM Press.

Ingwersen, P., & Järvelin, K. (2005). The Turn: Integration of information seeking and retrieval in context (The Information Retrieval Series). Secaucus, NJ: Springer-Verlag.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422–446.

Jiang, J., Han, S., & Lu, W. (2008). Expertise retrieval using search engine results. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008) Workshop on Future Challenges in Expertise Retrieval (fCHER) (pp. 11–16). New York: ACM Press.

Karimzadehgan, M., White, R.W., & Richardson, M. (2009). Enhancing expert finding using organizational hierarchies. In Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (pp. 177–188). Berlin, Heidelberg: Springer-Verlag.

Karimzadehgan, M., Zhai, C., & Belford, G. (2008). Multi-aspect expertise matching for review assignment. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (pp. 1113–1122). New York: ACM Press.

Kelly, D. (2006). Measuring online information seeking context. Part 1: Background and method. Journal of the American Society for Information Science and Technology, 57(13), 1729–1739.
Kim, J. (2009). Describing and predicting information-seeking behavior on the web. Journal of the American Society for Information Science and Technology, 60(4), 679–693.
Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.
Liebregts, R., & Bogers, T. (2009). Design and evaluation of a university-wide expert search engine. In Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval (pp. 587–594). New York: Springer-Verlag.
Menzel, H. (1960). Review of studies in the flow of information among scientists. New York: Columbia University, Bureau of Applied Social Research.
Metzger, M. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. Journal of the American Society for Information Science and Technology, 58(13), 2078–2091.
Rodrigues, S., Oliveira, J., & de Souza, J.M. (2005). Competence mining for team formation and virtual community recommendation. In Proceedings of the 9th International Conference on Computer Supported Cooperative Work in Design (Vol. 1, pp. 44–49).
Rosenberg, V. (1967). Factors affecting the preferences of industrial personnel for information gathering methods. Information Storage and Retrieval, 3(3), 119–127.
Serdyukov, P., & Hiemstra, D. (2008). Being omnipresent to be almighty: The importance of global web evidence for organizational expert finding. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008) Workshop on Future Challenges in Expertise Retrieval (fCHER) (pp. 17–24). New York: ACM Press.
Serdyukov, P., Rode, H., & Hiemstra, D. (2008). Modeling multi-step relevance propagation for expert finding. In Proceedings of the 17th ACM Conference on Information and Knowledge Management (pp. 1133–1142). New York: ACM.
Shami, S.N., Ehrlich, K., & Millen, D.R. (2008). Pick me!: Link selection in expertise search results. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2008) (pp. 1089–1092). New York: ACM Press.
Shami, S.N., Yuan, C.Y., Cosley, D., Xia, L., & Gay, G. (2007). That's what friends are for: Facilitating "who knows what" across group boundaries. In Proceedings of the ACM Conference on Supporting Group Work (GROUP 2007) (pp. 379–382). New York: ACM Press.
Soboroff, I., de Vries, A., & Crawell, N. (2007). Overview of the TREC 2006 Enterprise Track. In Proceedings of the 15th Text Retrieval Conference.
Terveen, L., & McDonald, D.W. (2005). Social matching: A framework and research agenda. ACM Transactions on Computer–Human Interaction, 12(3), 401–434.
Tsagkias, E., Larson, M., Weerkamp, W., & de Rijke, M. (2008). Podcred: A framework for analyzing podcast preference. In Proceedings of the Second Workshop on Information Credibility on the Web (pp. 67–74). New York: ACM Press.
Vakkari, P. (2001). Changes in search tactics and relevance judgements when preparing a research proposal: A summary of the findings of a longitudinal study. Information Retrieval, 4(3–4), 295–310.
W3C. (2005). The W3C test collection. Available at: http://research.microsoft.com/users/nickcr/w3c-summary.html
Weerkamp, W., & de Rijke, M. (2008). Credibility improves topical blog post retrieval. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT) (pp. 923–931). Stroudsburg, PA: Association for Computational Linguistics.
Wiig, K.M. (1997). Knowledge management: Where did it come from and where will it go? Expert Systems With Applications, 13(1), 1–14.
Woudstra, L.S.E., & Van den Hooff, B.J. (2008). Inside the source selection process: Selection criteria for human information sources. Information Processing & Management, 44, 1267–1278.
Yimam, D., & Kobsa, A. (2000). DEMOIR: A hybrid architecture for expertise modeling and recommender systems. In Proceedings of the Ninth IEEE International Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises (pp. 67–74). Gaithersburg, MD: NIST. Retrieved January 11, 2010, from http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00883706

## Appendix

*Questionnaire*

[*Introduction, Informed Consent, and space for answering questions are excluded from the example below, names are anonymized. The original questionnaire was made available in Dutch, the subjects' native language.*]

### Part I: Background Information

1. What is your job title?
2. What is your department/faculty?
3. In your daily work, how do you obtain information about research conducted in your faculty/department? Please name all sources of information that apply.
4. How often are you contacted with requests for experts?
5. Please give 1–2 examples of requests you have received.
6. How much time do you usually spend on answering such a request (including obtaining any information necessary and formulating a response)?
7. When you receive an expert request, how do you usually respond?
   - Contact the expert yourself
   - Forward the request to a suitable expert
   - Send the contact details of a suitable expert to the requesting party
   - Other: . . .

### Part II: Finding Similar Experts

[*This part of the questionnaire is printed on a new page and repeated for each expert*]

1. Assume you receive a request for an interview with Dr. Jane Doe. This person is on vacation. Whom do you recommend next?
2. Why would you recommend this person? Please name all factors that apply.
3. Would you consult any information sources in order to decide whom to recommend? If yes, which ones?
   [*The following is printed on the back of the page*]
4. Please consider the following list of experts. Please rank the experts according to how likely you would recommend each person, if Dr. Jane Doe was not available. Please assign each rank only once, from 10 (*most likely to recommend*) to 1 (*least likely to recommend*).

| | |
|---|---|
| Hubert Farnsworth | Seymour Skinner |
| Cary Granite | Arnie Pye |
| Turanga Leela | Milhouse Van Houten |
| John A. Zoidberg | Philip J. Fry |
| Alvin Brickrock | Selma Bouvier |

5. In the previous question, why did you choose the ranking as you did? Please describe all reasons that apply.

**Part III: Recommendation Factors**

[*Printed on a new page*]

When recommending experts, how much does each of the following factors influence your decision of whom you recommend?

| | Strong influence | Moderate influence | Low influence | No influence |
|---|---|---|---|---|
| Physical proximity | | | | |
| Availability | | | | |
| Approachability | | | | |
| Cognitive effort | | | | |
| Saves time | | | | |
| Topic of knowledge | | | | |
| Perspective | | | | |
| Reliability | | | | |
| Up-to-dateness | | | | |
| Familiarity | | | | |
| Sources contacts | | | | |

**Thank you very much for completing the questionnaire.**

Comments/Questions about this study: . . .


DOI: 10.1002/asi