On the Assessment of Expertise Profiles

Richard Berendsen and Maarten de Rijke
ISLA, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.
E-mail: {r.w.berendsen; derijke}@uva.nl

Krisztian Balog
Department of Electrical Engineering and Computer Science, University of Stavanger, NO-4036 Stavanger, Norway. E-mail: krisztian.balog@uis.no

Toine Bogers
Royal School of Library Information Science, Birketinget 6, DK-2300, Copenhagen, Denmark.
E-mail: tb@iva.dk

Antal van den Bosch
Faculty of Arts, CIW-Bedrijfscommunicatie, Radboud University Nijmegen, P.O. Box 9103, NL-6500 HD Nijmegen, The Netherlands. E-mail: a.vandenbosch@let.ru.nl

Expertise retrieval has attracted significant interest in the field of information retrieval. Expert finding has been studied extensively, with less attention going to the complementary task of expert profiling, that is, automatically identifying topics about which a person is knowledgeable. We describe a test collection for expert profiling in which expert users have self-selected their knowledge areas. Motivated by the sparseness of this set of knowledge areas, we report on an assessment experiment in which academic experts judge a profile that has been automatically generated by state-of-the-art expert-profiling algorithms; optionally, experts can indicate a level of expertise for relevant areas. Experts may also give feedback on the quality of the system-generated knowledge areas. We report on a content analysis of these comments and gain insights into what aspects of profiles matter to experts. We provide an error analysis of the system-generated profiles, identifying factors that help explain why certain experts may be harder to profile than others. We also analyze the impact on evaluating expert-profiling systems of using self-selected versus judged system-generated knowledge areas as ground truth; they rank systems somewhat differently but detect about the same amount of pairwise significant differences despite the fact that the judged system-generated assessments are more sparse.

Received April 23, 2012; revised December 4, 2012; accepted December 5, 2012

 DOI: 10.1002/asi.22908

Introduction

An organization's intranet provides a means for exchanging information and facilitating collaborations among employees. To efficiently and effectively achieve collaboration, it is necessary to provide search facilities that enable employees not only to access documents but also to identify expert colleagues (Hertzum & Pejtersen, 2006). At the Text REtrieval Conference Enterprise Track (Bailey, Craswell, de Vries, & Soboroff, 2008; Balog, Soboroff, et al., 2009; Craswell, de Vries, & Soboroff, 2006; Soboroff, de Vries, & Craswell, 2007), the need to study and understand *expertise retrieval* has been recognized through the introduction of the expert-finding task. The goal of *expert finding* is to identify a list of people who are knowledgeable about a given topic: *Who are the experts on topic X?* This task is usually addressed by uncovering associations between people and topics (Balog, Fang, de Rijke, Serdyukov, & Si, 2012); commonly, a co-occurrence of the name of a person with topics in the same context is assumed to be evidence of expertise. An alternative task, building on the same underlying principle of computing people–topic associations, is *expert profiling*, in which systems have to return a list of topics that a person is knowledgeable about (Balog, Bogers, Azzopardi, de Rijke, & van den Bosch, 2007; Balog & de Rijke, 2007). Essentially, (topical) expert profiling turns the expert-finding task around and asks the following: *What topic(s) does a person know about?*

Expert profiling is useful in its own right for users who want to profile experts they already know. It is also a key

task to address in any expert-finding system; such systems rank experts, and users will want to navigate to profiles of these experts. Complete and accurate expert profiles enable people and search engines to effectively and efficiently locate the most appropriate experts for an information need. In addition to a topical profile, it is recognized that social factors play a large role in decisions about which experts to approach (Balog & de Rijke, 2007; Cross, Parker, & Borgatti, 2002; Hofmann, Balog, Bogers, & de Rijke, 2010; Smirnova & Balog, 2011).

We focus on the topical expert-profiling task in a knowledge-intensive organization, that is, a university, and release an updated version of the Universiteit van Tilburg (UvT; Tilburg University [TU]) expert collection (Bogers & Balog, 2006), which was created with data from the UvT. Because the university no longer uses the acronym UvT and has switched to TU instead, we call the updated collection the *TU expert collection*.[1] The TU expert collection is based on the *Webwijs* ("Webwise") system[2] developed at TU. *Webwijs* is a publicly accessible database of TU employees who are involved in research or teaching, where each expert can indicate his or her skills by selecting expertise areas from a list of knowledge areas. Prior work has used these self-selected areas as ground truth for both expert-finding and expert-profiling tasks (Balog, 2008; Balog et al., 2007). With the TU expert collection come updated profiles consisting of these self-selected knowledge areas; we refer to this set of areas as *self-selected knowledge areas*.

One problem with self-selected knowledge areas is that they may be sparse. There is a large number of possible expertise areas to choose from (more than 2,000). When choosing their knowledge areas, experts may not necessarily browse the set of knowledge areas very thoroughly, especially because the interface in which they select the areas lists them in alphabetical order without providing links between related areas. This might result in sparse data with a limited number of knowledge areas assigned to each expert. Using these self-selected knowledge areas as ground truth for assessing automatic profiling systems may therefore not reflect the true predictive power of these systems. To find out more about how well these systems perform under real-world circumstances, we have asked TU employees to judge and comment on the profiles that have been automatically generated for them. Specifically, we have used state-of-the-art expertise retrieval methods to construct topical expertise profiles. TU employees were then asked to reassess their self-selected knowledge areas based on our recommendations; in addition, they were given the option to indicate the level of expertise for each selected area. Moreover, they could give free text comments on the quality of the expert profiles. We refer to this whole process as the *assessment experiment* in this article.

We group the research questions in this article in two parts. In the first part, we perform a detailed analysis of the outcomes of the assessment experiment. One important outcome is a new set of graded relevance assessments, which we call the *judged system-generated knowledge areas*. We examine the completeness of these new assessments. The knowledge areas experts selected and the textual feedback they gave provide us with a unique opportunity to answer the following question: "How well are we doing at the expert-profiling task?" We perform a detailed error analysis of the generated profiles and a content analysis of experts' feedback, leading to new insights on what aspects make expertise retrieval difficult for current systems.

In the second part, we take a step back and ask: "Does benchmarking a set of expertise retrieval systems with the judged system-generated profiles lead to different conclusions compared with benchmarking with the self-selected profiles?" We benchmark eight state-of-the-art expertise retrieval systems with both sets of ground truth and investigate changes in absolute system scores, system ranking, and the number of significant differences detected between systems. We find that there are differences in evaluation outcomes, and we are able to isolate factors that contribute to these differences. Based on our findings, we provide recommendations for researchers and practitioners who want to evaluate their own systems.

The main contributions of this article are as follows:

- The release of a test collection for assessing expert profiling—the TU expert collection—plus a critical assessment and analysis of this test collection. Test collections support the continuous evaluation and improvement of retrieval models by researchers and practitioners, in this case, in the field of expertise retrieval.
- Insights into the performance of current expert-profiling systems through an extensive error analysis, plus a content analysis of feedback of experts on the generated profiles. These insights lead to recommendations for improving expertise profiling systems.
- Insights into the differences in evaluation outcomes between evaluating the two sets of ground truth released with this article. This will allow researchers and practitioners in the field of expertise retrieval to understand the performance of their own systems better.

Before we delve in, we give a small recap of some terminology. Expert profiles, or topical profiles, in this article consist of a set of knowledge areas from a thesaurus. Throughout the article, we focus on two kinds of expert profiles that we use as ground truth.

Self-selected: These profiles consist of knowledge areas that experts originally selected from an alphabetical list of knowledge areas.

Judged system generated: These profiles consist of those knowledge areas that experts judged relevant from *system-generated profiles*: a ranked list of (up to) 100 knowledge areas that we generated for them.

[1]The TU expert collection is publicly available at http://ilps.science.uva.nl/tu-expert-collection. For a description of the items contained in the collection, please see the Appendix to this article.

[2]http://www.tilburguniversity.edu/webwijs/

DOI: 10.1002/asi

The rest of this article is structured as follows: We start by reviewing related work on test collection–based evaluation methodology in the Related Work section. In the Topical Profiling Task section, we define the topical profiling task. Next, we describe the assessment experiment: the profiling models used to generate the profiles and the assessment interface experts used to judge these profiles. In the Research Questions and Methodology section, we state our research questions and the methods used to answer them. We present and analyze the results of our assessment experiment in the Results and Analysis of the Assessment Experiment section, followed by an analysis of benchmarking differences between two sets of relevance assessments in the Self-Selected Versus Judged System-Generated Knowledge Areas: Impact on Evaluation Outcomes section. In the Discussion and Conclusion section, we wrap up with a discussion, conclusion, and look ahead.

Related Work

We start with a brief discussion on benchmarking and on how it has been analyzed in the literature. Then, we zoom in on the ingredients that constitute a test collection. Next, we consider related work on error analysis. We end with an overview of other test collections for expert profiling and expert finding.

Benchmarking

A recent overview on test collection–based evaluation of information retrieval systems can be found in Sanderson (2010). Today's dominant way of performing test collection–based evaluation in information retrieval was first carried out in the Cranfield experiments and later in numerous Text REtrieval Conference (TREC) campaigns. Our work falls into this tradition. To be able to create a yardstick for benchmarking, simplified assumptions have to be made about users, their tasks, and their notion of relevance. For example, in the TREC ad hoc collections, a user is assumed to have an information need that is informational (Broder, 2002), and a document is relevant if it contains a relevant piece of information, even if it is duplicate information. By framing the topical profiling task as a ranking task, we also make some simplifying assumptions. For example, users are satisfied with a ranking of expertise areas for an expert, and an area is relevant if experts have judged it so themselves.

Typically, evaluation methodologies are assessed by comparing them with each other, performing detailed analyses in terms of sensitivity, stability, and robustness. Sensitivity of an evaluation methodology has been tested by comparing it with a hypothesized correct ranking of systems (Hofmann, Whiteson, & de Rijke, 2011; Radlinski & Craswell, 2010). Stability and robustness are closely related concepts. An evaluation methodology can be said to be stable with respect to some changing variable, or robust to changes in that variable. For example, Radlinski and Craswell (2010) examine how evaluation changes when queries are subsampled. Buckley and Voorhees (2004) examine changes when relevance assessments are subsampled. In this study, we are interested in comparing and analyzing the outcomes of evaluating with two sets of relevance assessments—self-selected versus judged system-generated knowledge areas—and we consider two criteria: stability and sensitivity. To analyze stability, we identify four differences between our two sets of ground truth and ask how evaluation outcomes vary with respect to these differences. For analyzing sensitivity, we do not have a hypothesized correct or preferred ranking. Instead, we investigate how many significant differences can be detected with each set of ground truth.

When two evaluation approaches generate quantitative output for multiple systems, they can be correlated to each other. Often this is done by comparing the ordering of all pairs of systems in one ranking with the ordering of the corresponding pair in the other ranking. One often used measure is accuracy: the ratio of pairs for which both rankings agree (see, e.g., Hofmann, Whiteson, & de Rijke, 2011; Radlinski & Craswell, 2010; Sanderson & Zobel, 2005; Voorhees & Buckley, 2002). Another commonly used measure (Buckley & Voorhees, 2004; Voorhees, 2000) that we use in this article is Kendall tau (Kendall, 1938), a rank correlation coefficient that can be used to establish whether there is a monotonic relationship between two variables (Sheskin, 2011).

There are several ways to assess (relative) system performance besides benchmarking. Su (1992, 1994) directly interviews end users. Allan, Carterette, and Lewis (2005), Turpin and Scholer (2006), and Smith and Kantor (2008) give users a task and measure variables such as task accuracy, task completion time, or number of relevant documents retrieved in a fixed amount of time. Sometimes there are strong hypotheses about relative quality of systems by construction. Usage data such as clicks may also be used to estimate user preferences, for example, by interleaving the ranked lists of two rankers and recording clicks on the interleaved list (Hofmann et al., 2011; Joachims, 2002; Radlinski et al., 2008). In our study, we offer experts the opportunity to comment on the quality of system-generated profiles, which we analyze through a content analysis, as in Lazar, Feng, and Hochheiser (2010).

Ingredients of a Test Collection

A test collection–based evaluation methodology consists of a document collection, a set of test queries, a set of relevance assessments, an evaluation metric, and possibly a significance test to be able to claim that the differences found would generalize to a larger population of test queries.

Test queries. Test query creation in the TREC ad hoc tracks is typically done by assessors and, hence, test queries reflect assessors' interests. When test queries were created for the web track, they were retrofitted around queries sampled from web query logs so as to more closely reflect end-user interests (Voorhees & Harman, 2000). In our study on expert profiling, test queries (i.e., "information needs") are readily

DOI: 10.1002/asi

available: They are potentially all experts from the knowledge-intensive organization being considered.

At the TREC ad hoc tracks, test queries with too few or too many relevant documents are sometimes rejected (Harman, 1995; Voorhees & Harman, 2000). Harman (1995) reports that for all created queries, a trial run on a sample of documents from the complete collection yielded between 25 (narrow query) and 100 (broad query) relevant documents. Zobel (1998) notes that selecting queries based on the number of relevant documents may introduce a bias. In our experiments, we retain all test queries (i.e., all experts) that have at least one relevant knowledge area.

Relevance assessments. Relevance assessments are typically created by assessors. For the UvT collection used by Balog et al. (2007) and for the two new sets of relevance assessments we release with the TU collection, they are created by the experts themselves. Other test collections for expert-finding assessments have been provided by an external person (for the W3C collection [W3C, 2005]), or by colleagues (for the CERC [Bailey, Craswell, Soboroff, & de Vries, 2007] and UvT [Liebregts & Bogers, 2009] collections).

Voorhees (2000) studies the impact of obtaining relevance assessments from different assessors on system ranking using Kendall tau. Although different assessors may judge different documents relevant, system ranking is highly stable no matter from which assessor the assessments are taken. In our situation, the different sets of relevance assessments for each expert are created by the same expert but through different means: In one case, the assessor had to manually go through a list; in the other, the assessor was offered suggestions. We find that system ranking may be affected by these differences.

An important aspect is the *completeness* of relevance assessments. When test collections were still small, all items in it were judged for every test query (Sanderson, 2010). The experts who participated in our experiments have little time, however, and were simply not available to do this. A well-known method to evaluate systems without complete assessments is pooling. It was proposed by Jones and van Rijsbergen (1975) as a method for building larger test collections. The idea is to pool *independent* searches using any available information and device (Sanderson, 2010). In our study, we also perform a specific kind of pooling. We use eight systems to generate expertise profiles, that is, lists of knowledge areas characterizing the expertise of a person. The eight systems are not independent, but are all possible combinations of one of two retrieval models, one of two languages, and one of two strategies concerning the utilization of a thesaurus of knowledge areas. Unlike the methodology used at TREC (Voorhees & Harman, 2000), we do not take a fixed pooling depth for each run, perform a merge, and order results randomly for the assessor. Instead, to minimize the time required for experts to find relevant knowledge areas, we aim to produce the best possible ranking by using a combination algorithm. We also have something akin to a manual run: It consists of the knowledge areas selected by experts in the TU-*Webwijs* system. We confirm in this study that this manual run contributes many unique relevant results; that is, the automatic systems fail to find a significant amount of these knowledge areas.

Zobel (1998) performs a study on *pooling bias*. The concern here is that assessments are biased toward contributing runs and very different systems would receive a score that is too low. In particular, systems that are good at finding difficult documents would be penalized. For several TREC collections, Zobel found that when relevant documents contributed by any particular run were taken out, performance of that run would only slightly decrease. In our study, experts only judged those knowledge areas that the automatic systems found. We study the effect of regarding unpooled areas as nonrelevant on system ranking and find that it has hardly any impact. For this, like Buckley and Voorhees (2004), we use Kendall tau.

Significance tests. Significance tests are mostly used to estimate which findings about average system performance on a set of test queries will generalize to other queries from the same assumed underlying population of queries. A simple rule of thumb is that an absolute performance difference of less than 5% is not notable (Spärck Jones, 1974). Pairwise significance tests are common in cases when different systems can be evaluated on the same set of test queries. Voorhees and Buckley (2002) test the 5% rule of thumb with a fixed set of systems and a fixed document collection. Sanderson and Zobel (2005) extend this research and consider relative rather than absolute performance differences; they prefer a pairwise t test over the sign test and the Wilcoxon test. Smucker, Allan, and Carterette (2007) compared p values (for the null hypothesis that pairs of TREC runs do not differ) computed with five significance tests. They find that the Fisher pairwise randomization test, matched pairs Student's t test, and bootstrap test all agree with each other, whereas the Wilcoxon and sign tests disagree with these three and with each other. They recommend Fisher's pairwise randomization test, which is what we use.

In our work, we vary query sets and sets of relevance assessments. Then, keeping the significance test used fixed, we measure the average number of systems each system differs significantly from. We view this as a rough indication of the ability of each set of assessments to distinguish between systems. The difference in this number between sets of relevance assessments is a rough heuristic for the difference in sensitivity of the sets. Cohen (1995), who was interested in repeating a benchmarking experiment using a more stringent alpha value in the significance test, computed the average number of systems each system differs from for both values of alpha. He called the difference between these numbers the *criterion differential*, saying it is a rough heuristic for the difference in sensitivity of both alpha values.

DOI: 10.1002/asi

Error Analysis

Although test collections enable us to discriminate systems in their average performance over a set of queries with a certain reliability and sensitivity, Harman and Buckley (2009) stress that it is important to understand variance in performance over queries. Often, performance of single systems varies more over queries than performance on one query varies over systems. Variation in performance over queries does not simply correlate with the number of relevant documents; there is an interaction between query, system, and document collection (Voorhees & Harman, 1996). In the error analysis of our best performing system—the combining algorithm that was used to arrive at the set of judged system-generated knowledge areas—on the expert level, we find the same lack of correlation between number of relevant knowledge areas and system score. We are able to explain some of the performance differences between systems based on other properties of experts, however, such as their profession and the kinds of documents they are associated with in the collection. In addition to providing an analysis at the expert level, we provide one at the level of knowledge areas. We distinguish two categories: knowledge areas that are difficult to find and knowledge areas that are too often retrieved at high ranks ("false positives"). Related work in this area was done by Azzopardi and Vinay (2008), who define evaluation metrics that capture how well systems make individual documents accessible and point to interesting evaluation scenarios in which these metrics may be applied.

Other Test Collections for Expert Profiling and Expert Finding

The TU expert collection that we release is an update and extension of the collection released by Balog et al. (2007), which has previously been used for expert profiling and expert finding. To the best of our knowledge, no other test collections have been used for the expert-profiling task. Other test collections for expert finding include the W3C collection (W3C, 2005) and the CERC collection (Bailey et al., 2007). For these collections, relevance assessments were obtained manually, in different ways (cf. Ingredients of a Test Collection section). Automatic generation of test collections has also been done. Seo and Croft (2009) use Apple Discussions[3] forums as expertise areas and use the top 10 rated answerers for each forum as experts in the ground truth. Jurczyk and Agichtein (2007) consider the author ranking task, in which authors have to be ranked according to the quality of their contributions. This task is related to expert finding except that authors are not ranked by the quality of contributions on a specific query. They use Yahoo!Answers'[4] thumbs-up/thumbs-down votes and average number of stars received for best answers as ground truth for the author-ranking task.

[3]http://discussions.apple.com

[4]http://answers.yahoo.com

Topical Profiling Task

The TU expert collection is meant to help assess topical profiling systems in the setting of a multilingual intranet of a knowledge intensive organization. One can answer the question "What topics does an expert know about?" by returning a topical profile of that expert: a record of the types of areas of skills and knowledge of that individual and a level of proficiency in each (Balog et al., 2012). The task consists of the following two steps: (a) identifying possible knowledge areas and (b) assigning a score to each knowledge area (Balog & de Rijke, 2007). In an enterprise search environment, there often exists a list of knowledge areas in which an organization has expertise. In our test collection, this is indeed the case; therefore, we focus on the second step. We assume that a list of knowledge areas $\{a_1, \ldots, a_n\}$ is given and state the problem of assigning a score to each knowledge area (given an expert) as follows: What is the probability of a knowledge area (a) being part of the expert's (e) topical profile? We approach this task as one where we have to rank knowledge areas by this probability $P(a \mid e)$.

In the TU expert collection, for this task, systems receive the following ingredients as input:

- A query consisting of an expert ID (i.e., an organization-wide unique identifier for the person)
- A collection consisting of publications, supervised student theses, course descriptions, and research descriptions crawled from the *Webwijs* system of TU (All documents are either Dutch or English. The language is known for research and course descriptions, and is unknown for publications and student theses.)
- Explicit associations between the expert ID and documents in the corpus
- A thesaurus of knowledge areas (Knowledge areas are available in two languages: Dutch and English. All areas have a Dutch representation, for most of them an English translation is available as well.)

Given this input, the requested system output is a ranked list of knowledge areas from the thesaurus.

We note a small subtlety concerning the language of documents in the collection. In previous work (Balog, 2008), systems were evaluated on the subset of knowledge areas for which both a Dutch and an English translation were available; if an expert had selected a knowledge area without an English translation, for evaluation purposes, this knowledge area would be considered as nonrelevant. In this work, if an expert selects a knowledge area, we consider it as relevant, regardless of whether it has an English translation.

Assessment Experiment

We first describe the models we used to produce the *system-generated profiles*. Then we describe the assessment interface that experts used to judge these profiles.

DOI: 10.1002/asi

Automatically Generating Profiles

The system-generated profiles are created by combining the results of eight state-of-the-art expert-profiling systems in a straightforward way. In this subsection, we describe the eight systems and the combination method, and we list the parameter settings we use in this article. The eight expertise profiling systems that we use differ in three dimensions: First, two different retrieval models are used. Second, systems use either the Dutch or the English translations of the knowledge areas. Third, half of the systems treat knowledge areas as independent of each other, whereas the other half use a thesaurus of knowledge areas to capture the similarity between them. We briefly describe the models here.

The two retrieval models considered below take a generative probabilistic approach and rank knowledge areas a by the probability that they are generated by expert e: $P(a \mid e)$. In the first model, called Model 1 in Balog, Azzopardi, and de Rijke (2009), we construct a multinomial language model θ_e for each expert e over the vocabulary of terms from the documents associated with the expert. We model knowledge areas as bags of words, created from their textual labels (either Dutch or English). It is assumed that knowledge area terms t are sampled independently from this multinomial distribution, with replacement. Then, for Model 1, we have:

$$P(a \mid e) = P(a \mid \theta_e) = \prod_{t \in a} P(t \mid \theta_e)^{n(t,a)} \quad (1)$$

where $n(t,a)$ is the number of times term t occurs in a. In estimating $P(t \mid \theta_e)$, we apply smoothing using collection term probabilities, with unsupervised estimation of smoothing parameters. Specifically, we use Dirichlet smoothing and use the average representation length (i.e., the average number of terms associated with experts) as the smoothing parameter. In the second model, called Model 2 in Balog, Azzopardi, and de Rijke (2009), we estimate a language model θ_d for each document associated with an expert. Let this set of documents be D_e. We sum the probabilities of each of these documents generating the knowledge area. The terms in a are sampled independently from each document. Then, for Model 2, we have:

$$P(a \mid e) = \sum_{d \in D_e} P(a \mid \theta_d) = \sum_{d \in D_e} \prod_{t \in a} P(t \mid \theta_d)^{n(t,a)} \quad (2)$$

To estimate $P(t \mid \theta_d)$, we smooth using collection term probabilities as before, estimating smoothing parameters in an unsupervised way. As before, we use Dirichlet smoothing, but here we set the smoothing parameter to the average document length in the collection.

As for the language dimension, recall that knowledge areas come in two languages: $a = \{a_{\text{Dutch}}, a_{\text{English}}\}$. The Dutch retrieval models estimate $P(a_{\text{Dutch}} \mid e)$; the English systems estimate $P(a_{\text{English}} \mid e)$.

Systems that use the thesaurus rely on a similarity metric between a pair of knowledge areas, sim (a, a'). This similarity is taken to be the reciprocal of the length of the shortest path $\text{SP}(a, a')$ between a and a' in the thesaurus. If two knowledge areas are not connected, their similarity is set to zero. In addition, we use a parameter m for the maximal length of the shortest path for which we allow knowledge areas to have a nonzero probability. Formally,

$$\text{sim}(a, a') = \begin{cases} 1/\text{SP}(a, a'), & 0 < \text{SP}(a, a') \leq m \\ 0, & \textit{otherwise} \end{cases} \quad (3)$$

We describe the thesaurus graph in detail in the A Thesaurus of Expertise Areas section. Note that we do not distinguish between different types of relation in the graph and use all of them when searching for the shortest path. Next, we use $\text{sim}(a, a')$ to measure the likelihood of seeing knowledge area a given the presence of another knowledge area a':

$$P(a \mid a') = \frac{\text{sim}(a, a')}{\sum_{a''} \text{sim}(a'', a')} \quad (4)$$

The idea is that a knowledge area is more likely to be included in a person's expertise profile if the person is knowledgeable on related knowledge areas. This support from other knowledge areas is linearly interpolated with $P(a \mid e)$ using a parameter λ to obtain an updated probability estimate $P'(a \mid e)$:

$$P(a \mid e) = \lambda P(a \mid e) + (1 - \lambda) \left(\sum_{a'} P(a \mid a') P(a' \mid e) \right) \quad (5)$$

For all systems, once $P(a \mid e)$ has been estimated, we rank knowledge areas according to this probability and return the top 100 knowledge areas for a given user; because we only retrieve knowledge areas a where $P(a \mid e) > 0$, the result list may contain fewer than 100 items.

Merging systems' outputs. To arrive at the set of judged system-generated knowledge areas, we proceed as follows. We use the eight profiling systems just described (i.e., {Model 1, Model 2} × {Dutch, English} × {with thesaurus, without thesaurus}) to estimate the probabilities of knowledge areas for each expert. Let us denote this as $P_i(a \mid e)$ ($i = \{1, \ldots, 8\}$). These probabilities are then combined linearly to obtain a combined score $P(a \mid e)$:

$$P(a \mid e) = \sum_i \alpha_i p_i(a \mid e). \quad (6)$$

In addition, the top three knowledge areas retrieved by each profiling system receive an extra boost to ensure that they get judged. This is done by adding a sufficiently large constant C to $P(a \mid e)$.

Parameter settings. For the systems that use the thesaurus, we let $m = 3$ (Equation 3) and $\lambda = .6$ (Equation 5). For the combination algorithm (Equation 6), we let $\alpha_i = 1/8$ for all i. Furthermore, we set $C = 10$ and, again, we only retrieve knowledge areas for which $P(a \mid e) > 0$.

DOI: 10.1002/asi

Judging the Generated Profiles

The assessment interface used in the assessment experiment is shown in Figure 1. At the top of the page, instructions for the expert are given. In the middle, the expert can indicate the knowledge areas (called "Expertise areas" in the interface) regarded as relevant by ticking them. Immediately below the top 20 knowledge areas listed by default, the expert has the option to view and assess additional knowledge areas. The expert may or may not have examined all (up to 100) retrieved knowledge areas in the generated profile; this information was not recorded. System-generated knowledge areas that were in the original (self-selected) profile of the expert are pushed to the top of the list and are ticked by default in the interface, but the expert may deselect them, thereby judging them as nonrelevant. For the ticked knowledge areas, experts have the option to indicate a level of expertise. If they do not do this, we still include these knowledge areas in the graded self-assessments, with a level of expertise of three ("somewhere in the middle"). At the bottom of the interface, experts can leave any comments they might have on the generated profile.

Research Questions and Methodology

We organize our research questions into two subsections. The first subsection is concerned with the results of the assessment experiment. We study the completeness of the judgments gathered and the quality of the generated profiles; we answer these questions in the Results and Analysis of the Assessment Experiment section. The second subsection deals with the impact of using two sets of ground truth on evaluation outcomes; we answer these research questions in the Self-Selected Versus Judged System-Generated Knowledge Areas: Impact on Evaluation Outcomes section. Next, we briefly motivate each research question and outline the methods used to answer them.

Results and Analysis of the Assessment Experiment

The TU expert collection includes two sets of assessments: self-selected knowledge areas and judged system-generated knowledge areas. Our first research question concerns these two test sets of relevance assessments:

RQ1. Which of the two sets of ground truth is more complete?
Methods used: We construct the set of all knowledge areas that an expert judged relevant at some point in time, *either* by including it in the self-selected profile *or* by judging it relevant in the self-assessment interface. We then look at which of the sets of ground truth contains more of these knowledge areas.

Remember that the judged profiles were generated by a combination of state-of-the-art systems. Our next three research questions answer the following informal question: "How well are we doing?"

RQ2. What are the characteristics of "difficult" experts?
For example, does the number of relevant knowledge areas correlate with performance? Does the number of documents associated with an expert matter? Is there a significant difference between mean performance over different groups of experts, for example, PhD students versus professors?
Methods used: We look for correlations by visual inspection. We group experts by their position (job title) and look for significant performance differences using the Welch two-sample t test (Welch, 1947). Because we perform a number of comparisons we use an α value of $\alpha = .01$ to keep the overall Type I error under control.

RQ3. What are the characteristics of "difficult" knowledge areas?
Methods used: We identify knowledge areas that are often included in experts' self-selected profiles but are rarely retrieved in the system-generated profiles. In addition, we identify knowledge areas that are often retrieved in the top 10 ranks of system-generated profiles but never judged relevant by experts.

RQ4. What are important aspects in the feedback that experts gave on their system-generated profiles?
Methods used: In a content analysis, performed by two researchers, aspects are identified in a first pass over the data. In a second pass over the data, occurrences of these aspects are counted.

Self-Selected Versus Judged System-Generated Knowledge Areas: Impact on Evaluation Outcomes

Next, we analyze the differences in evaluation outcomes that arise when our two sets of relevance assessments are applied to assess expert-profiling systems. Our main research question is the following:

RQ5. Does using the set of judged system-generated knowledge areas lead to differences in system evaluation outcomes compared with using the self-selected knowledge areas?
When answering these questions, we consider four differences between the two sets of relevance assessments: (a) only a subset of experts has judged the system-generated knowledge areas, (b) self-selected knowledge areas that were not in the set of system-generated knowledge areas are considered nonrelevant in the judged system-generated profiles, (c) experts selected new knowledge areas from the system-generated profile, and (d) experts provided a level of expertise for most judged system-generated knowledge areas. We isolate the effect of each difference by constructing five sets of ground truth (self-selected profiles, judged system-generated profiles, and three intermediate ones), which we will detail later. We consider the effect of each difference on three dimensions; these are handled as separate subquestions.

RQ5a. How do the differences between the set of self-selected knowledge areas and the set of judged system-generated knowledge areas affect absolute system scores?
Methods used: We analyze nDCG@100 performance for each of the five sets of ground truth. nDCG@100 is a metric that rewards both high precision, high recall, and—in the case of graded relevance assessments—correct ordering of relevant knowledge areas.

RQ5b. How do the differences between the set of self-selected knowledge areas and the set of judged system-generated knowledge areas affect system ranking?
Methods used: We analyze differences in ranking with the five sets of ground truth. Following Voorhees (2000), we use Kendall tau. Like Sanderson and Soboroff (2007), we use the following formula:

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}} \tag{7}$$

DOI: 10.1002/asi

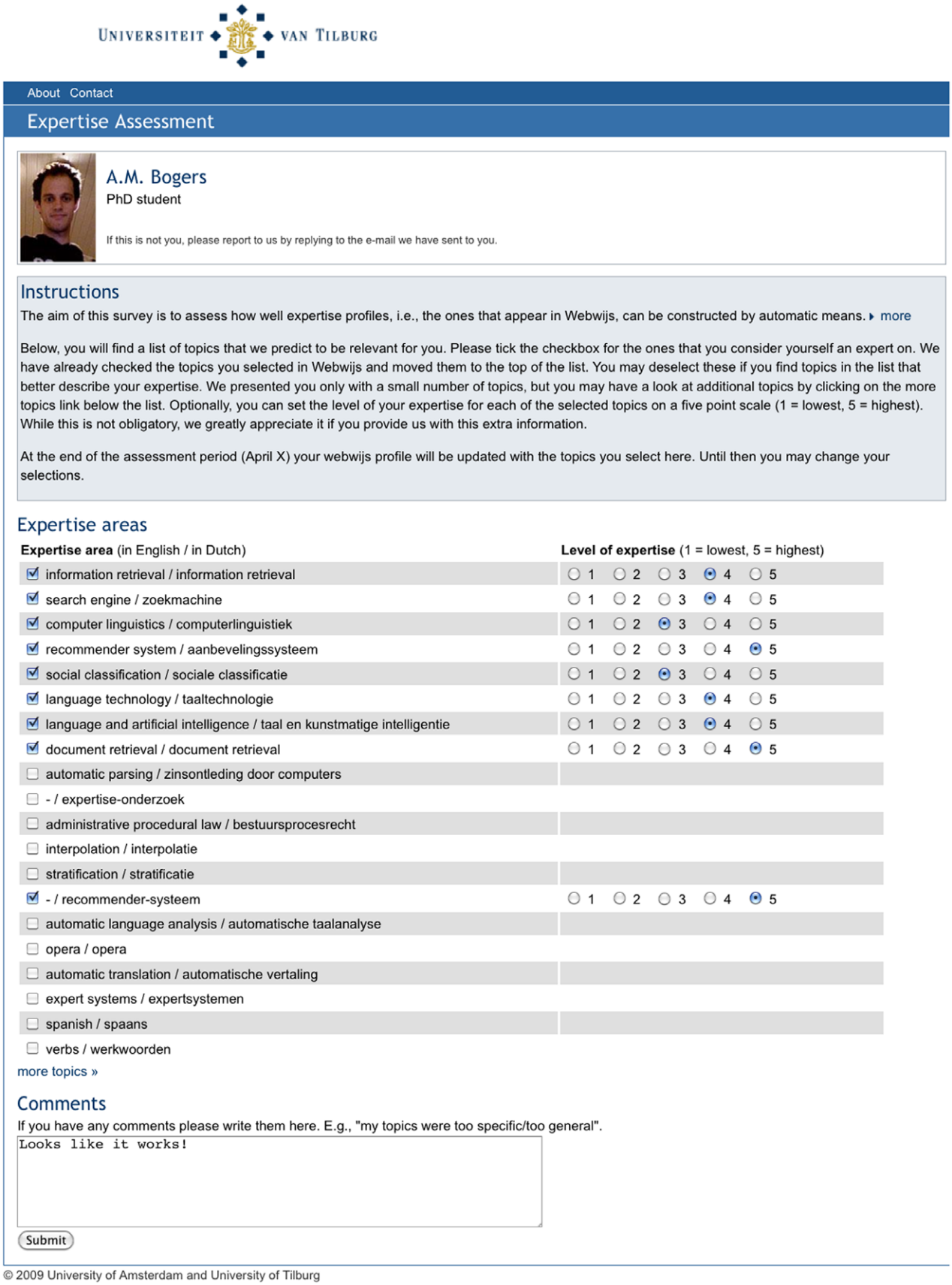

FIG. 1. Screenshot of the interface for judging system-generated knowledge areas. At the top, instructions for the expert are given. In the middle, the expert can select knowledge areas. For selected knowledge areas, a level of expertise may be indicated. At the bottom, there is a text field for any comments the expert might have. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

DOI: 10.1002/asi

where P is the number of concordant pairs, Q is the number of discordant pairs, T is the number of ties in the first list, and U is the number of ties in the second list. If there is a tie in at least one of the lists for a pair, the pair is neither correctly nor incorrectly ordered. When there are no ties, this formula is equivalent to the original formula as proposed by Kendall (1938). We compute Kendall tau for 28 pairs of system rankings. We accept a probability of Type I error $\alpha = .01$ for each comparison. Then the probability of at least one Type I error in all comparisons if they would be independent equals $1 - (1 - 0.01)^{28} = 0.25$. For eight systems, Kendall tau has to be greater than or equal to .79 to reject the null hypothesis. We do this analysis for four standard information retrieval evaluation metrics: mean average precision (MAP), mean reciprocal rank (MRR), normalized discounted cumulative gain calculated at depth 10 (nDCG@10), and nDCG@100. For MAP and MRR scores, trec_eval was used for evaluation; for implementing nDCG, we followed Clarke et al. (2008). We took all experts for the given test set into account during evaluation, even if systems did not retrieve any knowledge areas for them (these experts get zero score on all evaluation metrics).

RQ5c. How do the differences between the set of self-selected knowledge areas and the set of judged system-generated knowledge areas affect the average number of systems a system differs significantly from?

Methods used: We compare the five sets of ground truth on the basis of the number of significant differences in MAP, nDCG@100, MRR, and nDCG@10 that they detect between pairs of systems. A pair of systems differs significantly if their difference is expected to generalize to unseen queries. We use Fisher pairwise randomization test, following Smucker et al. (2007), and set $\alpha = .001$. We repeat this test for five sets of ground truth, four evaluation metrics (except that we have no MAP or MRR scores for the graded relevance assessments), and all possible $(\frac{1}{2} \cdot 8[8-1] = 28)$ pairs of systems: a total of 504 comparisons. Assuming that all of these comparisons are independent, this means accepting a Type I error of $1 - (1 - 0.001)^{504} = 0.40$. It is no problem for the interpretation of our results if there are a few spurious rejections of the null hypothesis; we mean to give an indication of the sensitivity of each set of ground truth, that is, the average number of systems that a system differs significantly from.

Results and Analysis of the Assessment Experiment

In this section, we report on the results of the assessment experiment defined in The Assessment Experiment section. We start with an examination of the completeness of the main tangible outcome of this experiment, the so-called judged system-generated knowledge areas. Then we analyze the quality of the generated profiles.

Completeness of the Two Sets of Ground Truth for Expert Profiling

To answer the question how complete each set of ground truth is (RQ1), we start out with some basic descriptive statistics. Our first set of ground truth contains 761 self-selected profiles of experts who are associated with at least one document in the collection. Together, these experts selected a total of 1,662 unique knowledge areas. On average, a self-selected profile contains 6.4 knowledge areas. The second set of ground truth contains 239 judged system-generated profiles. These experts together selected a total of 1,266 unique knowledge areas. On average, a judged system-generated profile contains 8.6 knowledge areas.

In Figure 2, the left two histograms show the distribution of experts over their number of relevant knowledge areas for the self-selected profiles (top) and for the judged system-generated profiles (bottom). The latter distribution is shifted to the right. The histograms on the right show the distribution of knowledge areas over the profiles that include them; the top right represents the self-selected profiles and the bottom right the judged system-generated profiles. The latter histogram is more skewed to the left; half of the knowledge areas have been judged relevant by a single expert only.

As an aside, we now check for how many of the graded judged system-generated knowledge areas we assigned our "somewhere in the middle" value of three, because the expert judged the knowledge area relevant without indicating a level of expertise. On average, this occurred for 0.6 of the 8.8 knowledge areas in each expert's profile. We conclude that the effect of this is negligible.

Now, to quantify the completeness of each set of ground truth in a single number, we proceed as follows. Let the set of all relevant knowledge areas associated with an expert be the union of the self-selected profile and the judged system-generated profile. Then subtract the knowledge areas that the expert deselected during the assessment interface (on average, experts removed 2% of the knowledge areas originally included in their self-selected profiles). We divide the resulting list of knowledge areas into three categories:

Only found by systems: These knowledge areas were not in the self-selected profile, but they were in the system-generated profile and were judged relevant by the experts.

Only found by experts: These knowledge areas were in the self-selected profile, but not in the system-generated profile.

Found by both: These knowledge areas were in both the self-selected and system-generated profiles, and the experts did not deselect them during the assessment experiment.

Table 1 lists the percentage of relevant knowledge areas that fall into each category, per profile, averaged over profiles. To answer RQ1, we find that the judged system-generated profiles are more complete. On average, a judged system-generated profile contains 81% (46% + 35%; see Table 1), whereas a self-selected profile contains only 65% (46% + 19%; see Table 1) of all relevant knowledge areas.

This leads to the following recommendation: Because the judged system-generated profiles are more complete, we expect this set of ground truth to give a more accurate picture of system performance, even if fewer assessed expert profiles are available. We elaborate on this when we answer RQ5 later in this article.

DOI: 10.1002/asi

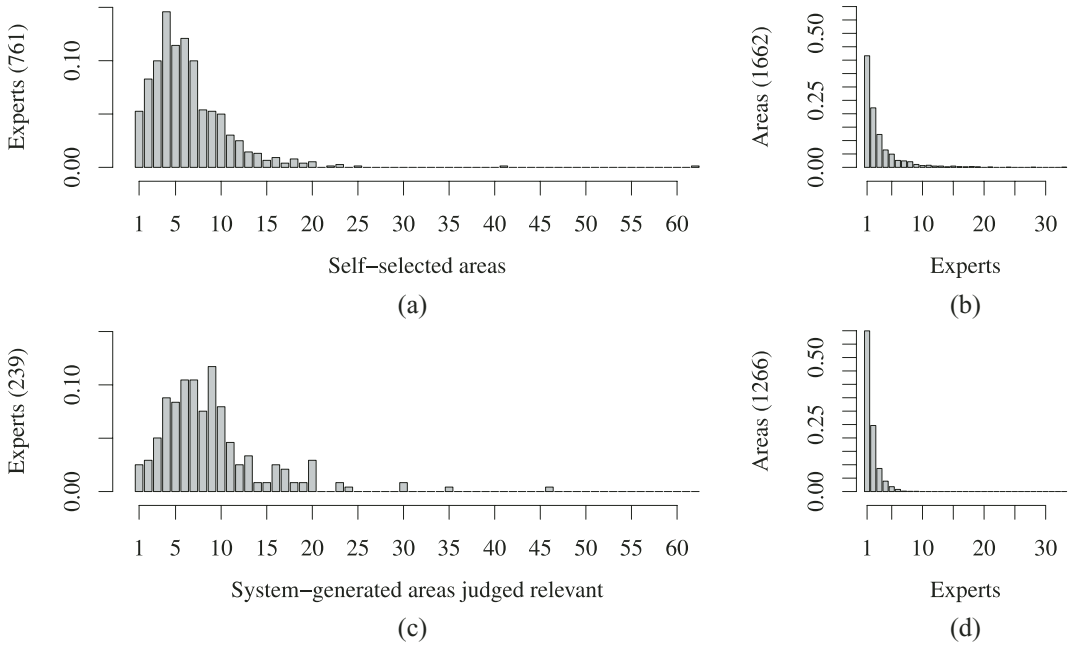

FIG. 2. Distribution of experts over their number of relevant knowledge areas (left) and distribution of knowledge areas over the profiles that include them (right). The top graphs are based on the self-selected profiles; the bottom graphs are based on the judged system-generated knowledge areas.

TABLE 1. Average percentage of the total number of relevant knowledge areas found for experts only by the automatic expert profilers, only by the experts when they self-selected knowledge areas, or by both.

	Average	Sample *SD**
Only found by systems	35%	24%
Only found by experts	19%	19%
Found by both	46%	24%

Note. * Sample *SD* over experts.

TABLE 2. Retrieval performance of the combined profiling algorithm on the self-selected and on the judged system-generated knowledge areas.

Ground truth	MAP	MRR	nDCG@10	nDCG@100
Self-selected (761)	0.16	0.40	0.21	0.36
Judged system generated (239)	0.43	0.71	0.44	0.66

Difficult Experts and Difficult Knowledge Areas

We investigate the characteristics of "difficult" experts (RQ2) and knowledge areas (RQ3). Before we begin with the analysis at the expert and knowledge area level, we report on the overall quality of the combined profiling algorithm in Table 2. We measure performance against the self-selected and the judged system-generated knowledge areas, respectively. All metrics are averaged over all profiles. Note that MAP and MRR treat relevance as a binary decision and the level of expertise indicated is not taken into account. Also note that there are no graded assessments available for the self-selected profiles; hence, nDCG@10 and nDCG@100 in the first row of Table 2 are computed using the same relevance level for all self-selected knowledge areas.

We find that considerably higher absolute scores are obtained on the judged system-generated profiles than on the self-selected ones. This finding holds for all metrics. Later, when we answer RQ5, we identify four factors that contribute to this large difference. In our detailed error analysis of the system-generated profiles that follows next, we focus on nDCG@100 because it is a metric that captures the quality of the entire system-generated profile.

Difficult experts. In this article, we aim to find properties of experts that can explain some of the variance in performance. We use the self-selected profiles of all 761 experts; this allows us to incorporate self-selected knowledge areas that were missing from the system-generated profiles in our analysis. We investigate a number of characteristics: the number of relevant knowledge areas for the expert, the number of documents associated with experts, and the position (job title) of an expert.

First, we attempt to find a correlation between these properties and nDCG@100 performance by visual inspection. We find no correlation between the number of relevant knowledge areas selected and nDCG@100, and no correlation between the number of documents associated with an expert and nDCG@100. Intuitively, the relationship between the ratio of relevant knowledge areas and number of documents associated with the expert is also interesting. For example, achieving high recall may be difficult when one has to find many knowledge areas in a few documents. Achieving high precision may be difficult if one has to find

DOI: 10.1002/asi

a few knowledge areas in many documents. However, we also find no correlation between the ratio of relevant knowledge areas and number of documents associated with an expert.

Next, we investigate a variable that may have different effects on performance indirectly: the position of an expert. In Figure 3, we see average nDCG@100 scores for the four most common positions among the 761 experts who self-selected a profile: lecturers (210), professors (168), PhD students (129), and senior lecturers (77); 99% confidence intervals on the estimated means are shown. These are calculated as $\bar{X} \pm 2.704 * \sigma / \sqrt{n}$, where σ is the sample *SD* and *n* is the sample size. The value 2.704 gives a 99% confidence interval for samples larger than 40. For professors, higher nDCG scores are achieved than for lecturers and PhD students; both of these differences are significant at the $\alpha = .01$ level (Welch two-sample *t* test).

An intuitive explanation for the fact that it seems easier to find relevant knowledge areas for professors than for PhD students is that professors have more publications. We just noted, however, that the number of documents associated with experts does not correlate with nDCG@100 performance. However, if we look a bit deeper into the different kinds of document that can be associated with an expert, we find that it matters whether an expert has a research description. Experts can have no research description, only a Dutch one, only an English one, or both a Dutch and an English one. We find that for the 282 experts without a research description, we achieve significantly lower average nDCG@100 performance than for the 479 experts who have at least one (Welch two-sample *t* test, $p < .001$). The difference, in absolute terms, is also substantial: .39 versus .30 for experts with and without a research description, respectively. It is not surprising that these research descriptions are important; they constitute a concise summary of a person's qualifications and expertise, written by the experts themselves. Of the professors, 73% have a research description against 53% of the PhD students, so this property explains part of the difference in performance between these two groups.

Missing knowledge areas. Next, we provide insights into relevant knowledge areas that we failed to retrieve in the system-generated profiles. To capture the fact that some knowledge areas are missing in more system-generated profiles than other knowledge areas, we define recall and precision measures for knowledge areas in a very straightforward and intuitive way. We say that knowledge areas that are missing in many system-generated profiles are *difficult*: They have low recall. Letting O_a be the set of self-selected profiles that contain knowledge area *a* and G_a, the set of system-generated profiles that contain *a*, we can define recall as follows:

$$R(a) = \frac{|O_a \cap G_a|}{|O_a|} \tag{8}$$

We are interested in knowledge areas *a* with low recall $R(a)$ here. Given equal recall, the more difficult knowledge areas are those that have lower precision:

$$P(a) = \frac{|O_a \cap G_a|}{|G_a|}. \tag{9}$$

We discard knowledge areas from our error analysis for which we cannot compute reliable recall and precision values. First, for computing recall, we exclude knowledge areas that are not in any self-selected profile. Also, we discard knowledge areas that are present in less than five self-selected profiles; the reason for doing so is to avoid large differences in recall for knowledge areas that may occur only by chance. Second, we cannot compute precision for knowledge areas that were not retrieved for any expert, which means only 14 (out of the 2,509) knowledge areas, 8 of which were also not in any self-selected profile. In this error analysis, therefore, we analyze only 361 of all 2,509 knowledge areas.

Figure 4 displays these 361 knowledge areas on a precision-recall plot. We added some jitter to the points for visualization purposes. In the bottom left corner of the figure, there are 17 knowledge areas with zero recall and precision. We list these "problematic" knowledge areas in Table 3, ranked by the number of system-generated profiles that contain them. This may be seen as an ordering by difficulty, where we consider knowledge areas that are more often retrieved incorrectly to be more difficult. In this list,

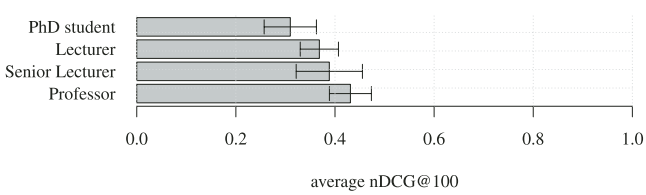

FIG. 3. Average nDCG@100 on the self-selected profiles for the four most common positions, with 99% confidence intervals.

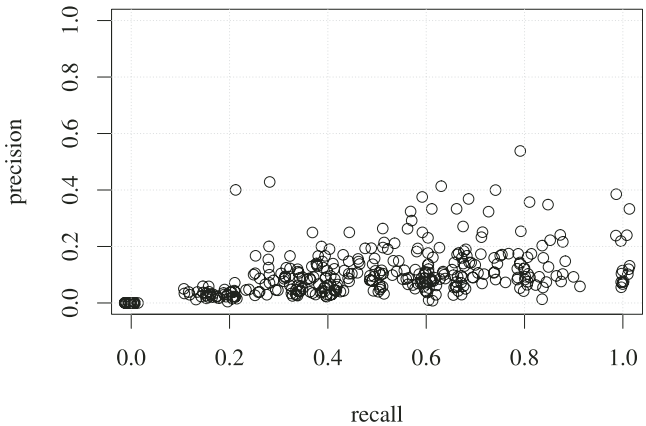

FIG. 4. Precision and recall of knowledge areas that were in at least five self-selected profiles.

DOI: 10.1002/asi

we find some very general knowledge areas such as *computer science* and *language*; there are also very specific knowledge areas such as *dutch for foreigners* and *income distribution*. Looking further down to the knowledge areas that are retrieved less often, we see many have no English translation. The English language profiling systems will never contribute these knowledge areas.

Knowledge areas often retrieved but never selected. We are interested in finding knowledge areas that are ranked high (e.g., in the top 10) for many experts and yet are always judged nonrelevant by these experts. For this analysis, we limit ourselves to the 239 system-generated profiles that have been judged in the assessment experiment.

In Figure 5, we show the distribution of knowledge areas over the number of experts they were retrieved for in the top 10.

Note that this distribution resembles the distribution of knowledge areas over the number of experts that judged them relevant in the assessment experiment (Figure 2b); this is a good property to have. We see that 1,395 knowledge areas are retrieved for at least one expert in the top 10; this is about 60% of all knowledge areas. Of these 1,395 knowledge areas, 773 were not judged relevant by any of the experts for whom they were retrieved in the top 10. We order these areas by decreasing number of system-generated profiles in which they were incorrectly included in the top 10, and show the top 20 in Table 4. Most of these knowledge areas appear to be quite specific.

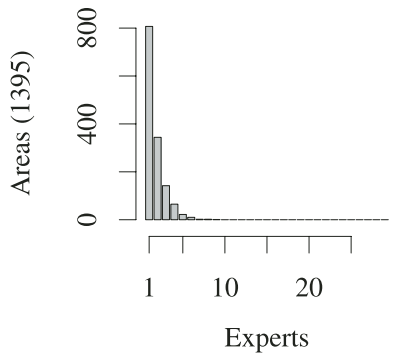

FIG. 5. Knowledge areas over the number of experts for whom they were retrieved in the top 10.

In summary, the main findings of this subsection are as follows. With regard to characteristics of difficult experts (RQ2): (a) Difficulty is *not* correlated simply with the number of relevant knowledge areas or with the number of documents associated with experts; (b) performance is significantly higher for experts who have a research description (in Dutch, English, or both). With regard to characteristics of difficult knowledge areas (RQ3), we find that knowledge areas that we often fail to retrieve (see Table 3): (a) often lack an English translation, making them impossible to find for our English-language profiling algorithms; and (b) can be both general and specific knowledge areas. Knowledge areas that we often retrieve in the top 10, although they were not judged relevant by experts (see Table 4), appear to be quite specific knowledge areas and sometimes lack an English translation.

Content Analysis of Expert Feedback

We now address our research question about important aspects in the feedback that experts gave by carrying out a content analysis (RQ4). During the assessment experiment, 91 experts left comments in the text area field at the bottom of the assessment interface. These comments were coded by two of the authors, based on a coding scheme developed by

TABLE 3. Problematic knowledge areas in terms of precision and recall.

Dutch	English	Missing	Retrieved	Added
informatica	computer science	8	71	0
inkomensverdeling	income distribution	5	66	1
taal	language	5	61	2
nederlands voor buitenlanders	dutch for foreigners	5	55	0
Automatisering	automation	5	49	1
culturele verscheidenheid	cultural diversity	5	41	2
e-government	e-government	6	39	0
evaluatie onderzoek	–	8	38	0
bedrijfsbeleid en -strategie	corporate policy and strategy	6	38	2
welzijn	well-being	6	26	1
ontwikkelingsvraagstukken	–	5	23	0
methoden en technieken, sociaal-wetenschappelijke	–	8	22	0
programmeren voor internet	–	7	20	0
beleidsonderzoek	–	8	18	1
cognitieve informatieverwerking	–	6	12	0
Kant, Immanuel (1724–1804)	Kant, Immanuel (1724–1804)	5	9	0
cultuurparticipatie	–	5	9	0

Note. For each knowledge area, we list the number of system-generated profiles where it is missing ("Missing") and where it is (incorrectly) retrieved ("Retrieved"). In a small number of cases, experts added these knowledge areas to their profile during the assessment experiment ("Added").

DOI: 10.1002/asi

TABLE 4. Knowledge areas shown in the top 10 but never selected.

Dutch	English	In top 10
alfabetisering in nederlands als tweede taal nt2	–	9
godsdienstpedagogiek	pedagogyofreligion	8
optietheorie	optionpricing	8
behendigheidsspelen	dexteritygames	7
sociolingustiek	sociolinguistics	7
asset liability management	assetliabilitymanagement	6
productiemanagement	productionmanagement	6
dienstenmarketing	servicesmarketing	6
werkloosheidsduur	–	6
mediarecht	–	6
cognitieve lingustiek	cognitivelinguistics	6
handelsmerken	trademarks	6
organisatiebewustzijn	–	6
belastingrecht	taxlaw	5
bestuursrecht	administrativelaw	5
geldwezen	money	5
instructieve teksten	instructivetexts	5
oefenrechtbank	mootcourt	5
onderwijs- en opleidingspsychologie	educationalandtrainingpsychology	5
openbaar bestuur	publicadministration	5

Note. The list is ordered by the number of times the knowledge area was retrieved in the top 10.

TABLE 5. Results of a content analysis of expert feedback.

Aspects	Count	Percentage
Quality of recommendations		
Excellent recommendations	7	7.9
Partially correct	10	11.2
All nonsense recommendations	15	16.9
Comments about individual knowledge areas		
Too much focused on one knowledge area	3	3.4
Missing key knowledge area (present in *Webwijs*, but not recommended)	32	36.0
Mix-up between different fields	2	2.2
One single nonsense recommendation	8	9.0
Comments about list as a whole		
Big overlap in recommended knowledge areas	10	11.2
Lack of consistency in recommended knowledge areas	1	1.1
Knowledge areas are too specific	4	4.5
Knowledge areas are too broad/general	10	11.2
No clear ordering of list	2	2.2
Upper limit of 10 knowledge areas	2	2.2
Knowledge areas taken from only one source (i.e., publications vs. theses)	5	5.6
Administrative comments		
Request to add expertise term to *Webwijs* itself	20	22.5
Missing *Webwijs* terms because of time difference between dump and survey	1	1.1
Complaint about incorrect or outdated *Webwijs* metadata	5	5.6
Rating expertise seen as ineffective	1	1.1
Complaint about spelling or translation of *Webwijs* knowledge areas	12	13.5

a first pass over these data. A statement could be assigned multiple aspects. After all aspect types were identified, the participants' comments were coded in a second pass over these data. Upon completion, the two annotators resolved differences through discussion. We report on interannotator agreement after discussion, reflecting cases where there remained a difference in opinion. We use two measures of interannotator agreement:

Micro-averaged interannotator agreement: The number of times both annotators coded a comment with the same aspect, divided by the total number of codings: 146/150 ≈ 0.97.

Macro-averaged interannotator agreement: For each aspect, interannotator agreement is calculated: the number of times both annotators coded a comment with this aspect, divided by the total number of codings with this aspect. Then the average of these aspect interannotator agreements is calculated: ≈ 0.98.

Both measures show very high interannotator agreement.

Table 5 lists all aspects with the count and the percentage of the comments in which they appeared.

First, we address the most common aspects in the experts' comments about the system-generated knowledge areas. The most common complaint is that a key knowledge area is missing. These missing knowledge areas were in *Webwijs*, and consequently, they were in the input list from which the retrieval systems select knowledge areas. This means the profiling algorithm is perceived to have insufficient recall. The second most frequently mentioned aspect is a request to add a new knowledge area to *Webwijs*. These do not reflect a failure on the part of our profiling algorithms. Rather, it is a request to the administrators of *Webwijs* to expand the thesaurus of knowledge areas. The third most common aspect is that the profile consists entirely of nonrelevant knowledge areas. This is a complaint about low precision. If there are relevant knowledge areas for the expert in the thesaurus, it also implies low recall.

Next, we examine the four categories of aspects in Table 5. Looking at the aspects relating to the quality of recommendations, we see that experts tend to be dissatisfied. We cannot directly relate this to average performance over all experts because we do not know the reasons why one expert chooses to leave a comment, whereas another decides not to. For some, dissatisfaction with the result list may be a motivation to comment, whereas others might find the results satisfactory enough and see no reason to add further feedback.

From comments about the lists as a whole, one of the main complaints is that there is much overlap in recommended knowledge areas. On one hand, this means that our algorithm finds "near misses," that is, knowledge areas that are not relevant but are very similar to relevant knowledge areas. On the other hand, it is clear that retrieving multiple very similar knowledge areas is not appreciated. De Rijke, Balog, Bogers, and van den Bosch (2010) propose a new metric that simultaneously rewards near misses and penalizes redundancy in a result list; we leave it as future work to actually implement and use this metric.

DOI: 10.1002/asi

A second main complaint about the list as a whole is that results are too general. Interestingly, the opposite complaint also occurs: Results are too specific. De Rijke et al. (2010) suggest that experts higher up in the organization tend to prefer more specific knowledge areas, whereas teachers and research assistants prefer broader terms. In our comments, complaints about a result list being too specific come from a professor, a lecturer, a researcher, and someone with an unknown function description. Complaints about the generated list being too general come from professors (5), senior lecturers (2), lecturers (2), and someone with no function description: mostly from senior staff.

In the administrative comments, it is interesting to note that almost no experts view rating knowledge areas as ineffective or unnecessary. Of course, experts were not explicitly asked about what they thought of rating knowledge areas, but still this was a big difference between our assessment interface and the *Webwijs* interface experts originally used for selecting knowledge areas.

To answer RQ4, the main aspects in the feedback of experts are (a) missing a key knowledge area in the generated profile (36%), (b) only nonrelevant knowledge areas in the profile (16.9%), (c) redundancy in the generated profiles (11.2%), and (d) knowledge areas being too general (11.2%).

In summary, it is clear that there is room for improvement in terms of both precision and recall. Because experts complain about redundancy in their profiles, in future work the diversity of profiles deserves attention. The desired level of specificity/generality is to a large extent a matter of personal preference. There are more complaints, however, about knowledge areas being too general; this is an indication that algorithms overall may score better by preferring specific knowledge areas.

Self-Selected Versus Judged System-Generated Knowledge Areas: Impact on Evaluation Outcomes

In this section, we look at another aspect of the TU expert collection as a measurement device. We study the differences between evaluating profiling systems with the self-selected knowledge areas and evaluating them with the judged system-generated knowledge areas (RQ5). The differences between the two types of assessment are isolated using five sets of ground truth, which we detail in the first subsection. In the remaining subsections, we study the changes between evaluating with the two types of assessment along three dimensions: absolute system scores (RQ5a), system ranking (RQ5b), and the average number of systems a system performs significantly different from (RQ5c). All of this is meant to help understand the merits of the TU expert collection.

Five Sets of Assessments

In the Results and Analysis of the Assessment Experiment section, we have studied the differences between self-selected and judged system-generated profiles; the corresponding ground truth that was used for evaluation (cf. Table 2) will be referred to as GT1 and GT5, respectively, throughout this section. These two sets of assessments differ on a number of dimensions: the number of profiles evaluated, the knowledge areas considered relevant within the profiles, and the grades of relevance. To help better understand the impact these differences might have on system evaluation, we introduce three more intermediate sets of assessments (GT2, GT3, GT4). Next, we briefly discuss each of the five sets.

GT1: Self-selected profiles. GT1 includes self-selected profiles of all experts for whom we generated a profile. Experts had previously selected these knowledge areas in the *Webwijs* system of TU; this set contains 761 experts.

GT2: Self-selected profiles of participants in assessment experiment. GT2 includes the self-selected profiles of only those experts who completed the assessment experiment. To be able to realize all subsequent evaluation conditions with the same set of experts, we limit this set of experts to the following groups:

- Those who completed the assessment experiment, selecting (or keeping) at least one knowledge area
- Those who had a nonempty self-selected profile
- Those for whom at least one of the knowledge areas in their self-selected profile was retrieved by the automatic profiling systems (This condition is required to be able to analyze, for the same set of experts, what evaluation differences there are when we evaluate only on the pooled subsets of their self-selected profiles.)

  As noted in the Completeness of the Two Sets of Ground Truth for Expert Profiling section, this set comprises 239 experts; for ease of reference, we sometimes refer to them as "our assessors."

GT3: Pooled subsets of self-selected profiles. For each self-selected profile of an assessor, we use only knowledge areas that were in the system-generated profile. This means that knowledge areas that are not in the system-generated profile are treated as nonrelevant.

GT4: Judged system-generated profiles (binary). GT4 includes the knowledge areas judged relevant during the assessment experiment. We consider only binary relevance. If a knowledge area was selected, it is considered as relevant; otherwise, it is taken to be nonrelevant.

GT5: Judged system-generated profiles (graded). GT5 is the same as GT4, but now with graded relevance. Experts could optionally indicate their level of expertise on each knowledge area they selected. Recall that when experts have selected a knowledge area but indicated no level, we assume they would have indicated a level "somewhere in the middle": level three out of five.

In the next subsection, we go through these five sets of ground truth, looking only at nDCG@100. We show how absolute system scores change from set to set.

Contrasting GT1 Through GT5

Previously, in our error analysis of system-generated profiles, we have seen that the combined profiling algorithm

DOI: 10.1002/asi

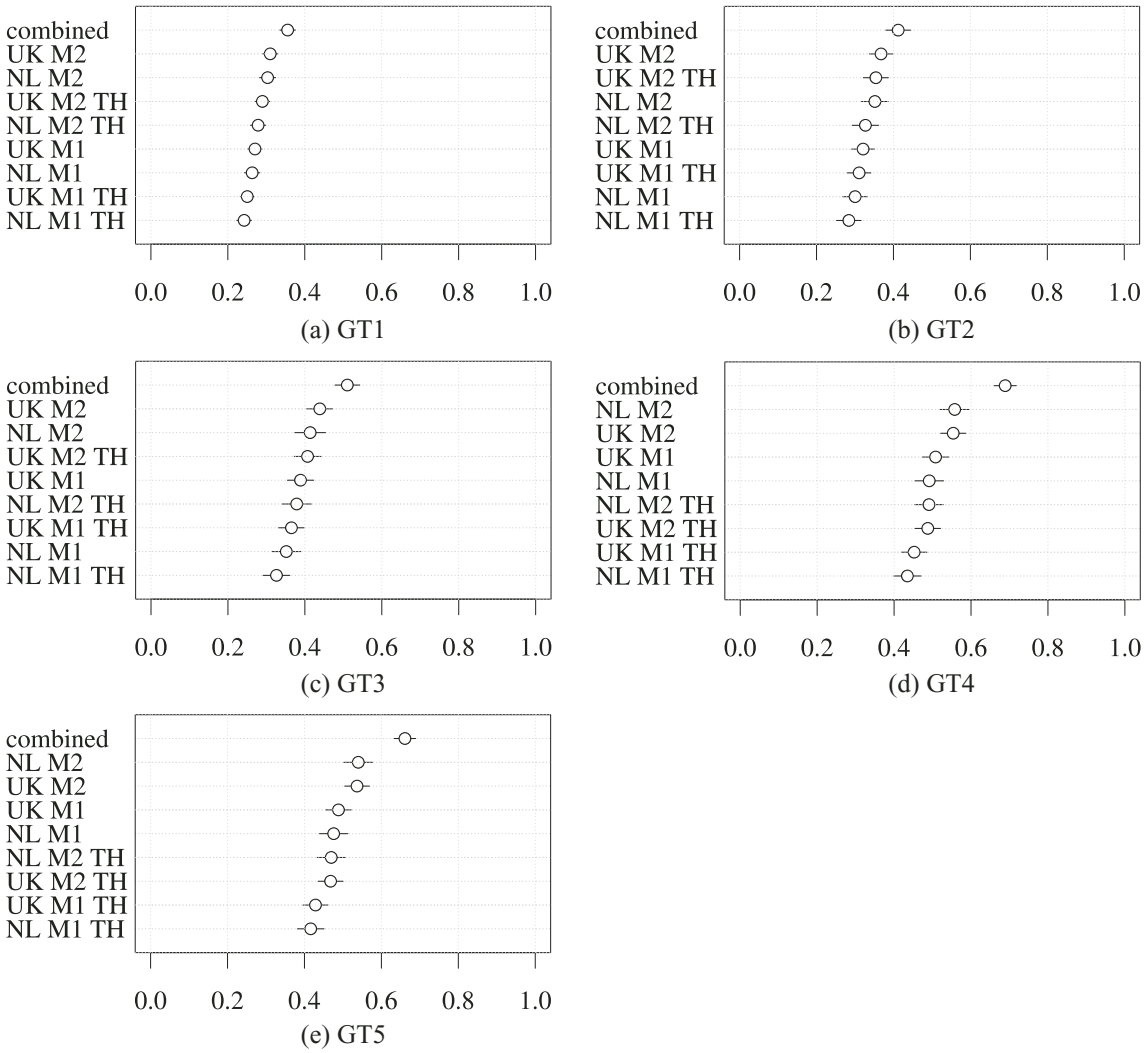

FIG. 6. Average nDCG@100 for each profiling system defined in The Assessment Experiment section, with 99% confidence intervals, for the five sets of assessments (GT1–GT5) examined in this section.

achieved higher scores on GT5 than on GT1. Here, we investigate the influence of the four differences between GT1 and GT5, in a step-by-step fashion, by considering each of GT1, . . ., GT5 and evaluating our profiling using those sets of assessments. To remain focused, we evaluate all systems with nDCG@100; nDCG is a well-understood metric that can be used with both binary and graded relevance assessments.

In Figure 6a, we show the nDCG@100 scores obtained with GT1 for all systems.[5] Confidence intervals on the means $\overline{X}$ are shown. These are based on the assumption that nDCG@100 scores are normally distributed. We show a 99% confidence interval, calculated as $\overline{X} \pm 2.576\hat{\sigma}/\sqrt{n}$, where $\hat{\sigma}$ is the sample *SD* and *n* is the sample size. The scores of individual systems are close to each other. Model 2 outperforms Model 1, the English-language systems tend to perform marginally better than their Dutch counterparts, and using the thesaurus does not appear to offer any benefits. The combined algorithm, which generated the profile shown to our assessors, outperforms all individual systems.

In Figure 6b, we plot the results based on GT2, on the self-selected profiles of assessors only. Confidence intervals are larger here; this is because the sample size is smaller. System scores are also a bit higher across the board.

In Figure 6c, we show the results obtained using GT3, that is, using only knowledge areas that the assessors may have seen during the assessment experiment. Recall that initially, experts see only the top 20 of the generated profile, but they can see up to 100 knowledge areas if they request more results. When we see the combined algorithm that generated the profile as a pooling algorithm, we can study the effect of pooling here. Absolute scores again increase for all systems. This is not surprising; unpooled knowledge areas are hard for all systems and regarding them as

[5]We use the following convention to name the profiling systems defined in The Assessment Experiment section: X Y Z, where X ∈ {NL, UK}, Y ∈ {M1, M2}, and Z ∈ { , TH}.

DOI: 10.1002/asi

nonrelevant reduces the problem difficulty. However, we can also see that relative system ranking hardly changes.

In Figure 6d, we evaluate profiling with GT4, that is, with the knowledge areas selected during the assessment experiment. We see a substantial performance increase in absolute scores for all systems, compared with evaluating with only the pooled knowledge areas from the original profiles. This increase is caused by knowledge areas that experts chose to *add* to their profiles. It is an indication that the original self-selected profiles were often incomplete, and systems are actually doing a better job than evaluating with the self-selected profiles would suggest. We also see changes in system rankings here that are a bit stronger than between other sets of ground truth. Systems that use Model 2 clearly outperform the ones that work with Model 1. Also, systems without the thesaurus are distinctly better than those with it. No language is preferred over the other.

For selected knowledge areas, experts could optionally indicate a level of expertise on a scale of one to five. In cases where they did not indicate a level of expertise for a selected knowledge area, we assigned a default level (3). In Figure 6e, we see how the move from binary to multiple levels of relevance changes nDCG@100: Absolute scores slightly decrease for all systems. This means that all systems retrieve knowledge areas in a suboptimal order. The relative ordering of systems, however, remains unchanged.

Answering RQ5a, we fine that: (a) Scores obtained on our assessors only are higher than on all experts. (b) Scores on only the pooled knowledge areas are higher than on the complete self-selected profiles; this is because self-selected knowledge areas that were unpooled are apparently difficult for these systems, and regarding them as nonrelevant reduces problem difficulty. (c) Scores on the binary judged system-generated knowledge areas are substantially higher than on the self-selected knowledge areas; this is an indication that the system-generated profiles were better than the original self-selected knowledge areas would give them credit for. (d) When we consider multiple relevance levels for assessment, absolute performance decreases a bit across the board, showing that to some extent all systems rank knowledge areas suboptimally.

Changes in System Ranking

We take a closer look at differences between GT1–GT5 and analyze whether and, if so, how they rank profiling systems differently. In the previous section we observed changes in system ranking in terms of nDCG@100, because of knowledge areas experts had added to their self-selected profile during the assessment experiment. In this section, we study how system rankings change on each set of GT1, . . ., GT5 for other metrics as well: MAP, MRR, and nDCG@10. We exclude the combined algorithm from our analysis here, because it produced the actual rankings that experts judged (experts are likely biased by the order in which suggestions were presented to them).

Tables 6 and 7 report Kendall tau for four evaluation metrics, computed between all pairs of sets of assessments. (The last rows of the tables are empty, because the MAP and MRR measures consider only binary relevance.) Table 6 shows the tau values for MAP and nDCG@100 in the lower and upper triangles, respectively. Both of these evaluation metrics capture precision as well as recall. Because all systems retrieve at most 100 documents, they both consider the complete list of results retrieved.

Let us consider the five sets of assessments GT1, . . ., GT5 for MAP and nDCG@10 and walk through Table 6. First, system ranking correlation between evaluating with the self-selected profiles of all 761 experts (GT1) and evaluating with the self-selected profiles of only the 239 assessors (GT2) is reasonable for both MAP and nDCG@100. Compared with GT2, considering unpooled knowledge areas as nonrelevant (GT3) ranks systems similarly for both metrics as well. Although we saw in the previous section that absolute scores increased substantially when unpooled knowledge areas are assumed to be nonrelevant, this has little effect on relative performance. The next step is including

TABLE 6. Kendall tau between system rankings on two sets of assessments with MAP (lower triangle) and average nDCG@100 (upper triangle).

	GT1	GT2	GT3	GT4	GT5
GT1 Self-selected profiles of all experts	–	0.86	0.86	0.57	0.57
GT2 Self-selected profiles of assessors	0.79	–	0.86	0.43	0.43
GT3 Pooled subsets of self-selected profiles	0.71	0.93	–	0.57	0.57
GT4 Judged system-generated profiles (binary)	0.57	0.50	0.57	–	1.00
GT5 Judged system-generated profiles (graded)	–	–	–	–	–

TABLE 7. Kendall tau between system rankings on two sets of assessments with MRR (lower triangle) and average n*DCG*@10 (upper triangle).

	GT1	GT2	GT3	GT4	GT5
GT1 Self-selected profiles of all experts	–	0.79	0.79	0.93	0.86
GT2 Self-selected profiles of assessors	0.71	–	1.00	0.71	0.79
GT3 Pooled subsets of self-selected profiles	0.71	1.00	–	0.71	0.79
GT4 Judged system-generated profiles (binary)	0.93	0.79	0.79	–	0.93
GT5 Judged system-generated profiles (graded)	–	–	–	–	–

DOI: 10.1002/asi

knowledge areas added during the assessment experiment (GT4). This does change the picture for both MAP and nDCG@100. For neither of the two metrics can we reject the null hypothesis, which states that there is no monotone relationship between the two rankings. Finally, taking into account the level of expertise (GT5) does not affect system ranking at all.

Next, we look at Table 7 and two measures that focus on the top ranks: MRR and nDCG@10. Table 7 shows Kendall tau values for MRR and nDCG@10 in the lower and upper triangles, respectively. Again, we step through the four changes that lead from GT1 to GT5. When evaluating with self-selected profiles from assessors only (GT2) instead of from all experts (GT1), rankings change a bit for both metrics. For MRR, there is no significant correlation. Regarding unpooled knowledge areas as nonrelevant (GT3) does not affect system ranking at all for these two metrics. Using the judged system-generated knowledge areas (GT4) instead of the self-selected knowledge areas changes the ranking a bit, again; for nDCG@10, there is no significant correlation. Finally, we find that the level of expertise (GT5) leads to only minor changes in system ranking for nDCG@10.

In answer to RQ5b, our findings are: (a) Comparing GT1 with GT2, the only difference being that GT2 evaluates with a subset of experts, we see that system rankings change a bit; nevertheless, for all metrics but nDCG@10, we can reject the null hypothesis, which states that system rankings do not correlate. (b) Regarding unpooled knowledge areas as nonrelevant hardly affects system rankings for the eight systems that contributed to the pool. Kendall tau values are high, ranging from 0.86 (nDCG@100) to 0.93 (MAP) to 1.00 (nDCG@10 and MRR) when comparing GT2 with GT3. (c) The knowledge areas that experts added to their self-selected profile during the assessment experiment have an effect on system rankings. When comparing GT1–GT3 with GT4, in all but two cases, we cannot reject the null hypothesis stating that there is no monotone relationship between system rankings obtained when evaluating with the self-selected versus judged system-generated profiles. (d) Comparing GT4 with GT5, we see that taking into account the level of expertise does not change system ranking for nDCG@10 or nDCG@100.

Pairwise Significant Differences

The final analysis we conduct concerns a high-level perspective: the sensitivity of our evaluation methodology. The measurement that serves as a rough estimate here is the average number of systems each system differs from; we compute this for each of the five sets of assessments and for four different metrics. We use Fisher's pairwise randomization test with $\alpha = .001$ to establish the average number of systems each system differs from in each condition. Table 8 lists these averages for MAP, MRR, nDCG@10, and nDCG@100. We start out with original profiles of all experts (GT1). If we limit ourselves to the 239 self-assessors (GT2), we see that the number of significant differences detected decreases for all four metrics. This is expected as the power of significance tests decreases with sample size. If we disregard nonpooled knowledge areas (GT3), we do not witness much change in the number of significant differences. Regarding nonpooled knowledge areas as nonrelevant does not change our insights about the relative performance of the profiling systems being examined. Comparing the self-selected profiles with the judged system-generated profiles (interpreted as binary judgments, GT4), there is a noticeable difference in the number of significant pairwise differences detected. For MAP and nDCG@10, we are roughly at the same level as for the self-selected profiles of all experts (GT1). If we use graded relevance (GT5), there is a slight increase for nDCG@10 and nDCG@100.

Answering RQ5c, we find that (a) fewer experts implies fewer significant differences, (b) regarding unpooled knowledge areas as nonrelevant does not have much effect on sensitivity, (c) knowledge areas that experts added to their profile during the assessment experiment lead to more detected significant differences, and (d) taking into account the level of relevance can lead to some further increase in sensitivity.

The two main findings for RQ5 overall are (a) GT4 (the judged system-generated knowledge areas, with binary relevance) is different from GT3, with much higher absolute scores, a different system ranking, and more detected pairwise significant differences between systems; and (b) for our eight systems, regarding the unpooled knowledge areas as nonrelevant does lead to higher absolute scores, but not to different system rankings or more detected pairwise significance differences.

Our findings lead to the following recommendations for researchers who would like to evaluate their expert-profiling systems on the TU expert collection. Because the judged system-generated profiles are more complete (see Completeness of the Two Sets of Ground Truth for Expert Profiling section), they form the preferred ground truth

TABLE 8. Average number of systems each system differs from significantly.

		MAP	MRR	nDCG@10	nDCG@100
Self-selected profiles	GT1 all experts	3.75	4.50	4.25	4.75
	GT2 assessors	2.75	2.25	2.75	3.00
	GT3 pooled subsets	2.75	2.25	3.25	2.75
Judged system-generated	GT4 binary	4.00	2.75	4.00	3.50
	GT5 graded	–	–	4.25	4.00

DOI: 10.1002/asi

for expert profiling. Compared with evaluating on the self-selected profiles, system ranking can change. Taking into account the level of expertise is useful because it does have an effect on absolute scores, even if it is not expected to lead to very different insights into relative system performance. If researchers are concerned that their methods are not rewarded for some retrieved knowledge areas that were not in the system-generated profiles, we recommend to repeat our analysis contrasting GT2 and GT3; this comparison allows for studying that factor in isolation.

Discussion and Conclusion

We released, described, and analyzed the TU expert collection for assessing automatic expert-profiling systems. The collection building process was detailed and we provided a critical assessment and analysis of this test collection. We started with an analysis of the completeness of self-selected versus judged system-generated knowledge areas as ground truth, an error analysis of system-generated expertise profiles, and a content analysis of feedback given by experts on system-generated expertise profiles. Then we took a step back and contrasted findings by benchmarking eight state-of-the-art expert-profiling systems with the two different sets of ground truth. We do not repeat all the answers to our research questions, but instead list the main findings for each, and with these main findings we give recommendations for the development and evaluation of expert-profiling systems. Then we discuss possible directions for future work for which the TU expert collection could be of use.

Main Findings With Recommendations

In this subsection, we repeat our research questions and list the main findings and recommendations.

RQ1. Which of the two sets of ground truth is more complete?
Judged system-generated profiles are more complete, on average. When we regard as relevant for an expert the union of knowledge areas in the self-selected profile and the judged system-generated profile (minus those knowledge areas that were judged nonrelevant), the average judged system-generated profile contains 81% and the self-selected profile 65% of all relevant knowledge areas.
Recommendation: It is preferable to use the system-generated profiles to evaluate expert-profiling systems because they are more complete.

RQ2. What are the characteristics of "difficult" experts?
Our main finding here is that experts who do not have a research description are significantly harder to profile accurately than experts who do.
Recommendation: Have experts in a knowledge-intensive organization maintain an up-to-date natural language description of their own expertise to facilitate better expert profiling.

RQ3. What are the characteristics of "difficult" knowledge areas?
Our main finding here is that knowledge areas that lack an English translation are more difficult to retrieve, and they are also among those knowledge areas that are most often retrieved without being relevant.
Recommendation: In a multilingual setting, maintain a complete translation of your list of knowledge areas in all languages to facilitate better expert profiling.

RQ4. What are important aspects in the feedback that experts gave on their system-generated profiles?
Experts mainly complain about missing a key knowledge area, generated profiles consisting of all nonsense knowledge areas, redundancy in the generated profiles, and retrieved knowledge areas being too general.
Recommendation: An interesting direction for future work is to go beyond ranking knowledge areas for an expert and to build coherent, complete, concise, diverse expertise profiles at the right level of specificity.

RQ5. Does using the set of judged system-generated knowledge areas lead to differences in system evaluation outcomes compared with using the self-selected knowledge areas?
We found that the knowledge areas experts added to their self-selected profile by judging them relevant do have an influence on system ranking, and we observe more significant differences between systems compared with evaluating with the self-selected profiles of these experts. Even though in the judged system-generated profiles some of the knowledge areas that experts had self-selected before are missing, this hardly affects the relative ranking of our eight systems.
Recommendation: It is preferable to use the judged system-generated profiles for benchmarking expert-profiling systems because these profiles are more complete. The missing knowledge areas from the self-selected knowledge areas hardly had an effect on relative performance from our systems, but if researchers wish to evaluate new and very different systems, we recommend to repeat our analysis contrasting the sets of ground truth GT2 and GT3 (we release all sets of ground truth used in this article).

Directions for Future Work

One conclusion that can be drawn from the error analysis and the content analysis of expert feedback is that there is still much room for improvement in the area of expertise retrieval. In addition to improving system performance on the task we studied in this article, we believe there are interesting possibilities to study tasks that differ subtly from it.

Expert profiling and expert finding. The expert-profiling task is closely related to the expert-finding task. Very similar algorithms may be used to approach the expert-finding and -profiling tasks; in both cases, the extent to which an expert and a knowledge area are associated have to be estimated (Balog et al., 2012). It has also been shown that expert-finding algorithms can benefit from the output of expert-profiling algorithms (Balog & de Rijke, 2007). In addition to benchmarking expert-profiling systems, the TU expert collection can also be used for benchmarking expert-finding systems. In this case, using the self-selected profiles would suit fine. Because the self-selected profiles are available for more experts, the number of relevant experts per knowledge area is somewhat larger in them. In addition, the graded relevance assessments were collected with the task of expert profiling in mind. Relevance levels are not guaranteed to be comparable across experts.

DOI: 10.1002/asi

Diversity, redundancy, and specificity. The evaluation metrics used in this article treat the relevance of knowledge areas in the ranked list independent from each other. We have seen that experts complained about redundancy in their generated profiles; something our evaluation metrics cannot capture. de Rijke et al. (2010) propose a metric that would reward diversity and near misses in a topical profile. Benchmarking with metrics like this is an interesting direction for future work. Experts complained about profiles being too general, and a few about profiles being too specific. One way to adjust the level of specificity in expertise profiles would be to require systems to organize knowledge areas in a hierarchy. A follow-up step could be to develop an assessment interface where experts can judge: (a) whether grouped knowledge areas are indeed similar, (b) whether hierarchical orderings are correct, and (c) whether retrieved knowledge areas are of the right specificity. The curated thesaurus that comes with the TU expert collection can be of help for work in this direction.

A learning assessment interface. The retrieval systems we evaluated in this article did not use knowledge areas that had already been self-selected by experts as evidence. This means our findings on their relative performance generalize to other settings where such self-selected ground truth is not available. Still, in settings where such ground truth is available, using it to locate additional relevant items is a powerful way of expanding a set of relevant items fast, with limited annotation effort. An assessment interface that would be fed by a learning retrieval model and would be continuously available for experts to update their profile is an interesting direction for future work.

Acknowledgments

We thank our reviewers for their many helpful comments and suggestions, and we thank Menno van Zaanen for providing up-to-date information about the *Webwijs* expertise selection interface. This research was partially supported by the IOP-MMI program of SenterNovem/The Dutch Ministry of Economic Affairs, as part of the À Propos project, the Radio Culture and Auditory Resources Infrastructure Project (LARM) as funded by the Danish National Research Infrastructures Program (project no. 09-067292), the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement 250430 (Galateas), the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreements 258191 (PROMISE Network of Excellence) and 288024 (LiMoSINe project), the Netherlands Organisation for Scientific Research (NWO) under projects 612.061.-814, 612.061.815, 640.004.802, 380-70-011, 727.011.005, 612.001.116, 277-70-004, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP and BILAND projects funded by the CLARIN-nl program, the Dutch national program COMMIT, and the ESF Research Network Program ELIAS.

References

Allan, J., Carterette, B., & Lewis, J. (2005). When will information retrieval be good enough? In 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 433–440). New York: ACM.

Azzopardi, L., & Vinay, V. (2008). Retrievability: An evaluation measure for higher order information access tasks. In 17th ACM Conference on Information and Knowledge Management (pp. 561–570). New York: ACM.

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

Bailey, P., Craswell, N., de Vries, A.P., & Soboroff, I. (2008). Overview of the TREC 2007 Enterprise Track. In The Sixteenth Text Retrieval Conference Proceedings (TREC 2007). NIST. Special Publication.

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

Balog, K., Azzopardi, L., & de Rijke, M. (2009). A language modeling framework for expert finding. Information Processing and Management, 45(1), 1–19.

Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., & van den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. In 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 551–558). New York: ACM.

Balog, K., & de Rijke, M. (2007). Determining expert profiles (with an application to expert finding). In Proceedings of the 20th international joint conference on Artificial intelligence (IJCAI07) (pp. 2657–2662). San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Balog, K., Fang, Y., de Rijke, M., Serdyukov, P., & Si, L. (2012). Expertise retrieval. Foundations and Trends in Information Retrieval, 6(2–3), 127–256.

Balog, K., Soboroff, I., Thomas, P., Craswell, N., de Vries, A.P., & Bailey, P. (2009). Overview of the TREC 2008 Enterprise Track. In The Seventeenth Text Retrieval Conference Proceedings (TREC 2008). NIST. Special Publication.

Bogers, T., & Balog, K. (2006). The UvT expert collection. Retrieved from http://ilk.uvt.nl/uvt-expert-collection/

Broder, A. (2002). A taxonomy of web search. ACM SIGIR Forum, 36(2), 3–10.

Buckley, C., & Voorhees, E. (2004). Retrieval evaluation with incomplete information. In 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 25–32). New York: ACM.

Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Bttcher, S., & MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In 31st annual international ACM SIGIR Conference on Research and development in information retrieval (pp. 659–666). New York: ACM.

Cohen, P. (1995). Empirical methods for artificial intelligence. Cambridge, MA: MIT Press.

Craswell, N., de Vries, A., & Soboroff, I. (2006). Overview of the TREC-2005 enterprise track. In The Fourteenth Text REtrieval Conference Proceedings (TREC 2005). NIST. Special Publication.

Cross, R., Parker, A., & Borgatti, S. (2002). A birds-eye view: Using social network analysis to improve knowledge creation and sharing. Knowledge Directions, 2(1), 48–61.

Harman, D. (1995). Overview of the second text retrieval conference (TREC-2). Information Processing & Management, 31(3), 271–289.

Harman, D., & Buckley, C. (2009). Overview of the reliable information access workshop. Information Retrieval, 12(6), 615–641.

Hertzum, M., & Pejtersen, A.M. (2006). The information-seeking practices of engineers: Searching for documents as well as for people. Information Processing & Management, 36(5), 761–778.

DOI: 10.1002/asi

Hofmann, K., Balog, K., Bogers, T., & de Rijke, M. (2010). Contextual factors for finding similar experts. Journal of the American Society for Information Science and Technology, 61(5), 994–1014.
Hofmann, K., Whiteson, S., & de Rijke, M. (2011). A probabilistic method for inferring preferences from clicks. In 20th ACM Conference on Information and Knowledge Management (pp. 249–258). New York: ACM.
Joachims, T. (2002). Evaluating retrieval performance using clickthrough data. In SIGIR Workshop on Mathematical/Formal Methods in Information Retrieval (pp. 12–15). New York: ACM.
Jones, K.S., & van Rijsbergen, C.J. (1975). Report on the need for and the provision of an ideal information retrieval test collection. British Library Research and Development Report, 5266, 43. University Computer Laboratory, Cambridge, United Kingdom.
Jurczyk, P., & Agichtein, E. (2007). Hits on question answer portals: exploration of link analysis for author ranking. In 30th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 845–846). New York: ACM.
Kendall, M. (1938). A new measure of rank correlation. Biometrika, 30(1/2), 81–93, 1938. Oxford University Press, London, United Kingdom.
Lazar, J., Feng, J., & Hochheiser, H. (2010). Research methods in human-computer interaction. New York: John Wiley & Sons Inc.
Liebregts, R., & Bogers, T. (2009). Design and implementation of a university-wide expert search engine. In 31st European Conference on Information Retrieval (pp. 587–594). Berlin, Heidelberg, Germany: Springer Verlag.
Radlinski, F., & Craswell, N. (2010). Comparing the sensitivity of information retrieval metrics. In 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 667–674). New York: ACM.
Radlinski, F., Kurup, M., & Joachims, T. (2008). How does clickthrough data reflect retrieval quality? In 17th ACM Conference on Information and Knowledge Management (pp. 43–52). New York: ACM.
de Rijke, M., Balog, K., Bogers, T., & van den Bosch, A. (2010). On the evaluation of entity profiles. In CLEF 2010: Conference on Multilingual and Multimodal Information Access Evaluation (pp. 94–99). Berlin, Heidelberg, Germany: Springer-Verlag.
Sanderson, M. (2010). Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, 4(4), 247–375. Now Publishers, Delft, The Netherlands.
Sanderson, M., & Soboroff, I. (2007). Problems with Kendall tau. In 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 839–840). New York: ACM.
Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In 28th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval (pp. 162–169). New York: ACM.
Seo, J., & Croft, W. (2009). Thread-based expert finding. In SIGIR 2009 Workshop on Search in Socal Media. New York: ACM.
Sheskin, D. (2011). Handbook of parametric and nonparametric statistical procedures. Boca Raton, FL: CRC Press.
Smirnova, E., & Balog, K. (2011). A user-oriented model for expert finding. In 33rd European Conference on Information Retrieval (pp. 580–592). Berlin, Heidelberg, Germany: Springer-Verlag.
Smith, C., & Kantor, P. (2008). User adaptation: Good results from poor systems. In 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 147–154). New York: ACM.
Smucker, M., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In 16th ACM conference on Information and Knowledge Management (pp. 623–632). New York: ACM.
Soboroff, I., de Vries, A., & Craswell, N. (2007). Overview of the TREC-2006 Enterprise Track. In The Fifteenth Text REtrieval Conference Proceedings (TREC 2006). NIST. Special Publication.
Spärck Jones, K. (1974). Automatic indexing. Journal of Documentation, 30(4), 393–432.
Su, L. (1992). Evaluation measures for interactive information retrieval. Information Processing & Management, 28(4), 503–516.
Su, L. (1994). The relevance of recall and precision in user evaluation. Journal of the American Society for Information Science, 45(3), 207–217.
Turpin, A., & Scholer, F. (2006). User performance versus precision measures for simple search tasks. In 29th Annual International ACM SIGIR conference on research and development in information retrieval (pp. 11–18). New York: ACM.
Voorhees, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management, 36(5), 697–716.
Voorhees, E., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 316–323). New York: ACM.
Voorhees, E., & Harman, D. (1996). Overview of the fifth Text REtrieval Conference (TREC-5). In The Fifth Text REtrieval Conference (TREC 1996). Special Publication.
Voorhees, E., & Harman, D. (2000). Overview of the ninth Text REtrieval Conference (TREC-9). In The Ninth Text REtrieval Conference (TREC 2000). Special Publication.
W3C. (2005). The W3C test collection. Retrieved from http://research.microsoft.com/users/nickcr/w3c-summary.html
Welch, B. (1947). The generalization of Students problem when several different population variances are involved. Biometrika, 34(1/2), 28–35.
Zobel, J. (1998). How reliable are the results of large-scale information retrieval experiments? In 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 307–314). New York: ACM.

Appendix

Description of the TU Expert Collection

The TU expert collection consists of a corpus of documents, a thesaurus of expertise areas, and two sets of relevance assessments. We devote a subsection to each. Before we start, we briefly introduce the original UvT collection and highlight the main differences to it.

Differences with the original UvT expert collection. The original UvT expert collection was harvested from the *Webwijs* ("Webwise") system developed at TU in the Netherlands. As explained in the Introduction of this article, *Webwijs* is a publicly accessible database of UvT employees who are involved in research or teaching. The UvT expert collection consists of four types of document: research descriptions, course descriptions, publications, and academic home pages. The majority of the data set was crawled in October 2006 (Bogers & Balog, 2006).

The TU expert collection was compiled in December 2008. The main reason necessitating the update is the fact that the data contained in the original version of the collection have become outdated; employees have left the organization, others have possibly changed their areas of interest, and new documents have been generated. One additional change we implement is to exclude academic home pages from the data set, because their usefulness has been found to be limited (Balog, 2008; Balog et al., 2007). Instead, we add another document type: *student theses*; these are bachelor's and master's theses of students supervised by researchers who are connected to TU.

Documents in the TU expert collection. Table A1 lists the types of documents available in the TU expert collection

DOI: 10.1002/asi

TABLE A1. Descriptive statistics of the TU expert collection.

Document type	Documents	People	People per document
Research descriptions (UK)	495	495	1.00
Research descriptions (NL)	524	524	1.00
Course descriptions	543	543	1.00
Publications	25,853	668	1.12
Student theses	5,152	520	1.17

Note. We list the number of documents of each type, the number of different people associated with documents of that type, and the average number of people associated with a single document.

along with some descriptive statistics. It is important to note about this new collection that XML files containing research descriptions include the previously selected expertise areas in the subject tags; this is ground truth. We did not index the contents of these tags, so this information is not exploited in our experiments. Researchers using the TU expert collection should take care to also disregard the contents of these subject tags if they want to benchmark expert-profiling systems.

Expertise areas in the TU expert collection. There is a total of 2,507 expertise areas. Each area is identified uniquely by a numeric identifier and has a Dutch textual label; most areas have an English translation as well. Expertise areas are organized in a thesaurus (see the following subsection).

A thesaurus of expertise areas. The thesaurus of expertise areas has broader-term/narrower-term relations between areas and related-term relations. In addition, it has preferred-term relations, where one area is the preferred term for another area. Of the total 2,507 knowledge areas, 2,266 are actually "approved."[6] When we discard areas that are not approved, there are 4,155 relations in the thesaurus; Table A2 lists the number of relations per type.

Using only the "broader term" relation, we can build a directed graph, with an edge pointing from area x to area y if x is a broader term than y. Apart from a few erroneous self-referencing areas, this graph is acyclic. Ignoring the direction of edges for a moment, we can find the connected components in this graph. There are no fewer than 635 connected components. One is very big with 718 edges; the second biggest has only 20 edges.

[6]New expertise areas can be suggested for inclusion in the thesaurus by TU employees. These suggested areas need to be reviewed by TU librarians and properly integrated into the thesaurus before they are fully approved.

TABLE A2. Binary relations between areas x and y in the thesaurus.

Abbreviation	Description	Count
BT	Area x is a broader term than area y	1,075
NT	The inverse of BT	1,076
USE	Area x is the preferred term for area y	247
UF	The inverse of USE	247
RT	Area x is related to area y	1,510

TABLE A3. Experts, total number of distinct relevant areas, and average number of areas per profile in both sets of ground truth.

	Experts	Areas	Average areas per profile
Self-selected profiles	761	1,662	6.4
Judged system-generated profiles	239	1,266	8.8

Two sets of relevance assessments. The TU expert collection comes with two main sets of relevance assessments, in the form of two files in standard trec_eval format. In both files, experts and areas are represented unique numeric identifiers. The first set of relevance assessments consists of profiles containing *self-selected areas* that experts selected from an alphabetic list of expertise areas. The second consists of profiles containing *judged system-generated areas*. Basic statistics about the number of experts and areas in both sets of ground truth are listed in Table A3. See also Figure 2 for the distribution of knowledge areas.

In addition to the two main sets of relevance assessments, we also release the three intermediate sets of relevance assessments used in the analysis in the Self-Selected Versus Judged System-Generated Knowledge Areas: Impact on Evaluation Outcomes section, so that researchers can repeat our analysis with their systems.

DOI: 10.1002/asi