Benchmarking Domain-Specific Expert Search Using Workshop Program Committees

Georgeta Bordea
Digital Enterprise Research Institute
National University of Ireland, Galway
name.surname@deri.org

Toine Bogers
Royal School of Library and Information Science
University of Copenhagen
Birketinget 6, 2300
Copenhagen, Denmark
tb@iva.dk

Paul Buitelaar
Digital Enterprise Research Institute
National University of Ireland, Galway
name.surname@deri.org

ABSTRACT

Traditionally, relevance assessments for expert search have been gathered through self-assessment or based on the opinions of co-workers. We introduce three benchmark datasets[1] for expert search that use conference workshops for relevance assessment. Our data sets cover entire research domains as opposed to single institutions. In addition, they provide a larger number of topic-person associations and allow a more objective and fine-grained evaluation of expertise than existing data sets do. We present and discuss baseline results for a language modelling and a topic-centric approach to expert search. We find that the topic-centric approach achieves the best results on domain-specific datasets.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]; H.3.1 [**Content analysis and indexing**]

Keywords

expertise search, expert finding, workshop dataset

1. INTRODUCTION

Leveraging the knowledge of informed people is essential in organisations and online communities as well as in scientific research, in a multitude of scenarios and settings. These include finding qualified reviewers to assess the quality of research submissions [12, 13], identifying consultants and collaborators inside or outside a community, locating topical experts for requests from the media [10], and discovering solvers in open innovation platforms [16], to name just a few. The most widely accepted approaches for measuring and analysing scientific research rely on publication metadata, focusing on publication counts or the number of citations. However, textual descriptions of scientific research, such as publication titles, abstracts and, increasingly, full-text content call for methods that allow a deeper content-based analysis of scientific output. Currently, content-based methods for analysing research expertise received more interest in an enterprise environment, where metadata about documents alone is not enough to identify experts on a specific topic. A successful expert search system helps address two important tasks, as signaled by Maybury [11]: *expert finding* and *expert profiling*. Expert finding is the task of locating individuals or communities knowledgeable about a specific topic, while expert profiling is the task of constructing a brief overview about the expertise topics of a person.

Evaluating expert search systems remains a challenge, however, despite a number of data sets that have been made publicly available in recent years [1, 2, 15]. Traditionally, relevance assessments for expert finding were gathered either through self-assessment or based on opinions of co-workers. Self-assessed expertise profiles are often subjective and incomplete and the opinions of colleagues are biased towards their social and geographical network. A recent study showed that, in an organisational setting, people are more likely to recommend as experts peers from their collaboration network or people that are geographically close [14].

We address these limitations by exploiting expertise data generated over the course of several decades in a peer-review setting, more specifically data about 590 conference workshops in Computer Science. Conference workshops are focused events organised around a narrow set of (interrelated) topics. Workshop organizers tend to be topical experts, as well as the program committee (PC) members that are invited to review submitted contributions to the workshop. A peer-review setting alleviates the problem of subjective assessments of expertise, as several community members have to reach an agreement. Typically, workshop proposals are subject to careful scrutiny before being accepted as part of a conference. Whether the organisers and the PC members are well-recognised experts in their field is an important criteria for the acceptance of a workshop proposal. The process of composing program committees for workshops is particularly interesting, as organisers have a broad knowledge of the domain and of domain experts. We limit ourselves to data about workshops, as conferences are more broad in scope than a workshop, making the assignment of topics to specific committee members increasingly difficult.

Another limitation of existing datasets for expert search is the sparsity of topic-person associations. This is partially due to the fact that relevance assessments are gathered through heavily involved and time-consuming interviews with experts. Instead, calls for workshop papers (CfPs) are more easy to collect and readily provide rich descriptions of areas of interest, that can be associated

[1]The datasets are available at http://itlab.dbit.dk/~toine/?page_id=631

CompSci'13, October 28, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2508497.2508501 ...$15.00.

with organisers and PC members. Hence, workshop CfPs are a valuable source of topic-expert associations, providing detailed descriptions of areas of interest and extensive lists of domain experts. This information can be used as a gold standard for expert finding, but only as a silver standard for expert profiling as the profiles inferred from workshop topics are incomplete. It is unlikely that a researcher will be invited as a workshop PC member in each of the areas that they are an expert in.

Previous test collections for expert search have typically focused on searching a single organization for experts and expertise. Different types of organizations were considered, such as research institutes [1, 15] and universities [2, 14]. Our test collection is focused around entire research communities in specific research domains, with members that are geographically dispersed and that have similar interests to some extent. In particular, we gathered data produced by two communities with a long tradition in Computer Science, Information Retrieval (IR) and Computational Linguistics (CL), as well as the more recent Semantic Web (SW) community. This requires methods that distinguish expertise at more fine-grained levels. Take for example the university use case, where expertise topics as broad as mathematics, philosophy, or physics are informative enough to distinguish between experts from different departments. When analysing expertise in a research community such as the IR or the CL community, more fine-grained expertise topics are required.

Our contributions in this paper are two-fold:

- We introduce three new test collections for expert search focused on entire research domains instead of single organisations. The properties of these test collections enable both research into expert search as well as scientometrics.
- We present state-of-the-art performance runs on these test collections to use as benchmarks for future work using these test collections.

The remainder of this paper is organized as follows. We start by presenting a review of related work in Section 2 and we introduce three datasets for Expert Search in Section 3. The Expert Search baselines are presented in Section 4, followed by the experimental results in Section 5. We conclude in Section 6, giving a few directions for future work.

2. RELATED WORK

Early large-scale approaches to expert finding involved the manual construction and querying of databases containing representations of the knowledge and skills of an organization's workforce, placing the burden and responsibility of maintaining them on the employees themselves [11]. This disadvantage prompted a shift to more automated expert finding techniques that supported the natural expertise location process [6].

A lot of ground was covered in terms of evaluating expert search systems by the organisation of three consecutive enterprise tracks by TREC [1], that provided common ground for evaluating different systems and approaches. Two enterprise datasets were made available to the community: a 2004 crawl of the World Wide Web Consortium (W3C) website used in the 2005 and 2006 editions of the TREC Enterprise track [15], and a dataset from the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) [1]. Both datasets contain hundreds of thousands of documents contributed by thousands of people, but they provide only a small number of topics that are either identified by task organisers or collected from people inside the organisation. Furthermore, document-candidate associations are not explicitly provided, and have to be estimated based on candidate occurrences in text. The UvT Expert Collection [2] addresses some of these limitations by introducing a more realistic number of topical areas, 1500 compared to only about 50 for the previous two collections, and by unambiguously identifying document authors. Associations between topics and experts are self-selected for increased realism, but relatively sparse with an average of 6 topics assigned per expert. An updated version of the UvT collection was described in more detail by Berendsen *et al.* (2013) [3].

In recent years, other approaches to expert search using different data sets have been presented as well, although the data sets used in these approaches have not been all made publicly available. For example, the DBLP Computer Science Bibliography, a computer science bibliography website, has been used as a resource in expert search and scientometrics before. Tang *et al.* (2008) performed the extraction and mining social networks on DBLP data using their ArnetMiner system [17]. They propose a probabilistic framework for author name disambiguation, use topic modeling to match authors, venues, and documents to topics, and extracted the citation network between the publications in their DBLP crawl. In addition, they released their 2008 crawl of DBLP augmented with this citation information, which we use as the basis for our IR test collection as described in Section 3.1.

Deng *et al.* (2008) also used a crawl of DBLP to evaluate their expert finding algorithms: a language modeling approach weighted by the strength of the expert-document associations, and a topic modeling approach [7]. They evaluate their algorithms on the entire DBLP data set, i.e., all domains as opposed to just the IR-focused part we describe in Section 3.1. To alleviate the problem of sparse textual content in DBLP, they take the title of a publication and aggregate it with the top ten Google Scholar results when using that title as a query. While this does reduce textual sparsity, it is also likely to introduce noise in the document representations. They used university researchers to assess the relevance of expert finding results against short key phrases that represent general expertise areas as opposed to the more fine-grained expertise areas our collections focus on. In addition, they did not make their data set publicly available.

A related problem that shares many similarities with expert finding as we approach it in this paper is the automatic assignment of conference and journal submissions to reviewers. Typically, these approaches use the sets of papers written by the individual reviewers as content-based expertise evidence for those reviewers to match them to submitted papers [4, 8, 9, 18]. The most extensive work was done by Yarowksy *et al.* (1999), who performed their experiments on the papers submitted to the ACL '99 conference [18]. They compared both content-based and citation-based evidence for allocating reviewers and found that combining both sources of information resulted in the best performance. Our second test collection—described in more detail in Section 3.2—covers the same domain, but it is much larger, covering the entire ACL Anthology Reference Corpus.

3. DATA COLLECTIONS

We present three new test collections in this paper, that are focused around entire research communities in a specific research domain as opposed to just a single organization. While this section provides some general considerations about the datasets, sections 3.1-3.3 describe some of the particularities of each dataset. For each test collection, we describe the process of collecting the documents, creating the topics and producing the relevance assessments. Table 1 contains an overview of the main characteristics of our three test collections, including information about the total

number of documents, workshops, and authors from each research area.

Each dataset consists of a corpus of documents and information about their authors along with a collection of workshops from the same research areas that can be used as basis for a gold standard evaluation. The CL and the SW datasets are collected and maintained by the research communities that initially published the works. Therefore, relatively clean metadata about events and publications, including full-text content, are directly available from the same location. The same cannot be said about the IR collection that has to be gathered from different sources including DBLP, Google Scholar, and ArnetMiner [17]. This resulted in a smaller coverage of full-text publications as can be seen in Table 1. Organizers and PC members correspond naturally to (a subset of) the relevant experts on the topic of the workshop. To generate a topic description for each workshop, we extracted the title of the workshop as well as a *short* and a *long* description of the purpose of the workshop. The long description of the workshop was typically taken from the starting page and covered the complete description of the goals and focus of the workshop (except for the areas of interest). The short description typically corresponded to the first paragraph of the long description: one or two sentences containing a concise teaser description of the workshop[2]. In addition, we extracted the *areas of interest* from each workshop website, which are often presented as a bullet-point list of research areas. Extracting this workshop information was done by a group of annotators.

It is often the case that a workshop is organised with the intent to create a platform for communication between research areas with overlapping interests. For example the workshop on the emergent topic of "Computational Neurolinguistics" organised by the CL community in 2010 was meant to bring together researchers from the areas of computational linguistics and cognitive neuroscience that have an interest in machine learning methods. Such workshops have PC members with expertise backgrounds that match one of these areas, or a combination of both. To allow a more fine-grained identification of experts we manually annotated each workshop from the IR, CL and SW areas with expertise topics. The list of organizers and PC members for each workshop served as our relevance assessments: people listed for a workshop were considered to be relevant experts for the topic of the workshop. To represent the likely difference in expertise between organizers of a workshop and PC members, we assigned a relevance value of 2 to the organizers and a value of 1 to the PC members. Five different annotators were involved in constructing the topic set for the IR collection; for the CL and SW collections, 2 annotators were involved.

The datasets proposed in this work can be used as a basis for the investigation of several different tasks. The CL and the SW datasets provide a large number of documents and a list of associated terms extracted from workshop descriptions that can be used for *term extraction* (1) or *expertise topic extraction* (2) evaluation. The main difference between these two tasks is that generally expertise topics have to be broad enough to summarize a knowledge area, while terms can be more specific. Another task that can be addressed is the *assignment of experts to program committees* (3) using workshop descriptions. In this way workshop organizers can identify PC members based on their interests mentioned in previous publications. *Expert finding* (4) is a similar task that takes as input more focused keyphrase-based descriptions of topics instead of a workshop description. Finally, extracted terms can be assigned to topical profiles, a task that is known as *expert profiling* (5). Profiling a candidate requires the identification of areas of skills and knowledge that best describe their interests and expertise.

[2]Distinguishing between what constitutes a short and long version of the description was left up to the individual annotator; we did not check for inter-annotator agreement, although incidental inspection suggested a consistent extraction process.

Table 1: Overview of our three test collections (IR = Information Retrieval, CL = Computational Linguistics, SW = Semantic Web).

	IR	CL	SW
#documents	24,690	10,921	2,311
% of full-text documents	54.1%	100%	100%
#workshops	60	340	190
#unique authors	26,098	9,983	4,480
#authors/document	2.7	2.2	3.3
#experts/workshop	14.9	25.8	24.9
#expertise topics	488	4,660	6,751

3.1 Information Retrieval

The first set of research domains covers the related fields of information retrieval (IR), digital libraries (DL), and recommender systems (RS). To construct a test collection covering all of these research fields, we used the DBLP Computer Science Bibliography[3], a computer science bibliography website that tracks the most important journals and conference proceedings in computer science. Our initial motivations for constructing a test collection around DBLP were two-fold: (1) the fields of IR, DL, and RS are well-covered in DBLP, and (2) a special version of the DBLP data set, augmented with citation information, is available from the team behind ArnetMiner, which allows for investigations into the use of citation information for expert search.

Topics & relevance judgments To make the augmented DBLP collection suited to expert search evaluation, we needed realistic topic descriptions as relevance judgments at the expert level. To collect these, we turned to workshops organized at the major conferences covering the fields of IR, DL, and RS. To identify relevant workshops, we visited the websites of the CIKM, ECDL, ECIR, IIiX, JCDL, RecSys, SIGIR, TPDL, WSDM, and WWW conferences, which have substantial portions of their program dedicated to IR, DL, and RS. We collected links to workshop websites for all workshops organized at those conferences between 2001 and 2012. This resulted in a list of 60 different workshops with websites that were still online at the time of writing[4].

Document collection To construct our test collection covering the aforementioned fields, we took the augmented DBLP data set released by the team behind ArnetMiner as our starting point. This data set is a October 2010 crawl of of the DBLP data set containing 1,632,442 different papers with 2,327,450 citation relationships between papers in the data set[5]. As this augmented data set contains publications from all fields of computer science, we filtered out all publications not belonging to IR, DL, and RS by restricting ourselves to publications in relevant journals, conferences, and workshops.

We created this list of relevant venues in two steps. First, we generated a list of *core venues* by extracting all papers published

[3]Available at http://dblp.uni-trier.de/, last accessed July 9, 2013.

[4]The list of 60 active workshops can be viewed at http://itlab.dbit.dk/~toine/?page_id=631.

[5]Available at http://arnetminer.org/DBLP_Citation, last accessed July 9, 2013.

at the conferences used for topic creation: CIKM, ECDL, ECIR, IIiX, JCDL, RecSys, SIGIR, TPDL, WSDM, and WWW. We select these conferences, because as hosts to the topic workshops, they are likely to be relevant venues for the PC members to publish in. This resulted in a data set containing 9,046 different publications from these core venues. However, restricting ourselves to these venues alone means we could be missing out on experts that tend to publish more in journals and workshops. We therefore extended the list of core venues with other venues tracked by DBLP that also have substantial portions of their program dedicated to IR, DL, and RS. Venues that only feature incidental overlap with IR, such as the Semantic Web conference, were not included. We also excluded venues that did not have 5 publications or more in the augmented DBLP data set. While this does exclude the occasional on-topic publication in venues that are pre-dominantly about other topics, we believe that this strategy will cover the majority of relevant publications. This additional filtering step resulted in a final list of 78 *curated venues* (core plus additional)[6] covering a total of 24,690 publications.

In addition to citation information, the augmented DBLP data set was also extended with abstracts wherever available. However, the team behind ArnetMiner was only able to add abstracts for 33.7% of the 1.6 million publications (and 43.5% of the 24,690 publications in our test collection). We therefore attempted to download the full-text versions of al 24,690 publications using Google Scholar. We constructed a search query consisting of the last name of the first author and the full title without surrounding quotes[7]. We then extracted the download link from the top result returned by Google Scholar (if available). We were able to find downloads URLs for 14,823 of the 24,690 publications in our filtered DBLP data set for a recall of 60.04%, where recall is defined as the percentage of papers in our filtered DBLP data set that we could find download URLs for. While this is not as high as we would like, it does represent a substantial improvement over the percentage of abstracts present in the augmented DBLP data set. Moreover, a recall rate of 100% is impossible to achieve as tutorials, keynote abstracts, and even entire proceedings are typically not available online in full-text, but they are present in the DBLP data set.

Around 90.15% of these download URLs we obtained in this manner were functional, which means we were able to download full-text publication files for 13,363 publications (or 54.12% of our entire curated data set). We performed a check of 100 randomly selected full-text files to see if these were indeed the publications we were looking for and achieved a precision of 97% on this sample. We therefore assume that the false positive rate of our approach is acceptably low.

3.2 Computational Linguistics

The ACL Anthology Reference Corpus[8] is a second dataset made available by the Computational Linguistics community. Documents are collected from events such as ACL, EACL, NAACL, SemEval, ANLP, EMNLP, Coling, HLT, IJCNLP, and LREC. Workshop descriptions are easily identified in the directory structure of the dataset,as they are grouped together under the same folder, and organised based on the year when the event was held. Each workshop is associated with a document that describes the event. One annotator manually extracted information about the organizers, PC members, year, and title from these documents. At the same time, each workshop was annotated with terms that describe the main areas of interest. For example, the workshop on "Biomedical Information Extraction" from 2009 was annotated with terms such as: biomedical information extraction, biomedicine, health care, healthcare delivery, personalized medicine, and clinical narrative. All of these terms were explicitly mentioned in the description of the workshop.

3.3 Semantic Web

The third dataset is a corpus of scientific publications from Semantic Web conferences[9] that were published in the proceedings of several conferences, including: ISWC, EKAW, ESWC, WWW, ASWC, and I-Semantics [10]. The dataset is available through a public SPARQL endpoint. Workshops along with information about title, year, description, website, organizers, and PC members, can be queried by selecting all the events of type *WorkshopEvent*. In some cases this information is missing or incomplete and we used the provided workshop website to manually extract the data. We were not able to do this for a subset of workshops that did not have active websites anymore. A large number of terms that appear in workshop descriptions and that describe the main topics of interest were manually annotated. For example the workshop with the tile "1st International Workshop on Stream Reasoning" from 2009 was annotated with terms including stream reasoning, reasoning, network monitoring, data streams, traffic engineering, and sensor networks. These terms are associated with each PC member and workshop organiser and can be used as relevance assessments for expert finding and expert profiling.

4. EXPERT SEARCH BASELINES

In this section we describe two baselines for Expert Search and we discuss how they can be applied for expert finding and expert profiling in a peer-review setting.

Language modelling baseline

The first baseline is based on generative language modelling (LM) and was proposed in [2]. The problem of finding experts for a given query is formulated in terms of computing the probability $p(ca|q)$ of a candidate ca being an expert given a query topic q. In this setting, both the expert finding and the expert profiling tasks depend on accurate estimations of the probability $p(q|ca)$. This probability can be modelled in several ways. A first approach is to construct a language model for each candidate by aggregating the information from all the documents that they authored. This corresponds to method *LM1* in our experiments. The alternative is to build a language model for documents, find the most relevant documents for a given query and then check who authored those documents. This methods is referred to as *LM2* in Section 5. For both approaches we used the open source implementation available online[11].

Topic-centric baseline

A topic-centric (TC) baseline puts emphasis on the extraction of keyphrases that can succinctly describe expertise areas, also called expertise topics, using term extraction techniques [5]. This method is referred to as *TC* in Section 5. An advantage of a topic-centric approach is that topical profiles can be constructed directly from text, without the need for controlled vocabularies or manual identification of terms.

[6]The list of curated active workshops is available at http://itlab.dbit.dk/~toine/?page_id=631.

[7]A preliminary test on just the publications from the core venues showed that adding quotes around the publication title decreased recall from 80.3% to 70.86%.

[8]Available at http://acl-arc.comp.nus.edu.sg/, last accessed July 15, 2013.

[9]Available at http://data.semanticweb.org, last accessed July 15, 2013.

[10]The complete list of conferences can be found here: http://data.semanticweb.org/conference

[11]Ears toolkit: http://code.google.com/p/ears/

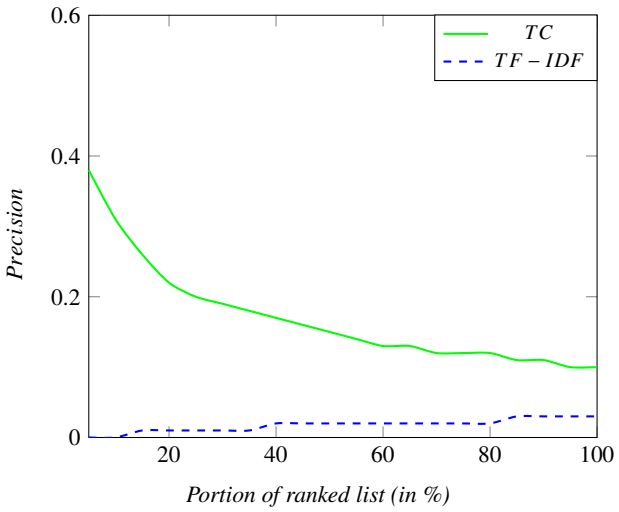

Figure 1: Precision for top 10k terms from the CL workshops dataset

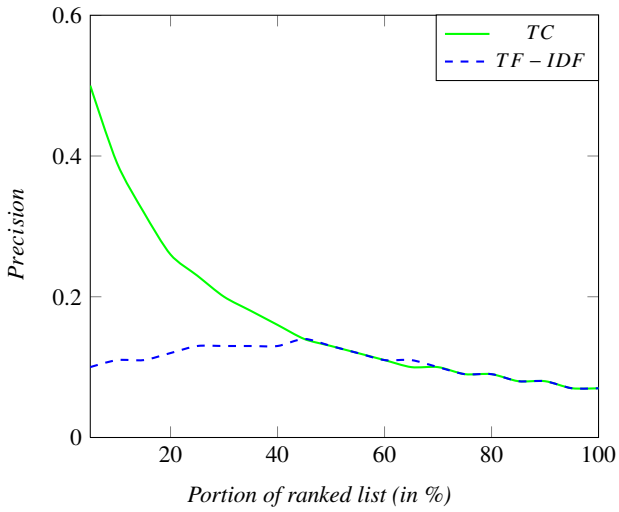

Figure 2: Precision for top 10k terms from the SW workshops dataset

Table 2: Expert finding results for the language modelling approach (LM), document count(DC), and the topic centric approach (TC)

Dataset	Measure	LM1	LM2	DC	TC
CL	MAP	0.0071	0.0056	0.0324	**0.0327**
	MRR	0.0631	0.0562	0.2648	**0.2650**
	P@5	0.0202	0.0173	**0.1292**	0.1277
SW	MAP	0.0070	0.0067	**0.0317**	0.0288
	MRR	0.0528	0.0522	**0.2235**	0.1963
	P@5	0.0182	0.0188	**0.1030**	0.0921
IR	MAP	0.0599	0.0402	0.1345	**0.1407**
	MRR	0.1454	0.1231	0.3829	**0.3851**
	P@5	0.0614	0.0485	0.1609	**0.1644**

Expertise topic extraction is implemented as follows. First, candidate expertise topics are discovered from text using a syntactic description for terms (i.e., nouns or noun phrases) and some contextual patterns that insure that the candidates are coherent within the domain. Because the test collection introduced in this paper is from Computer Science, we were able to use the ACM Computing Classification System[12] to manually identify a list of 80 words that are representative for the domain. These domain-specific words were used to select as candidates noun phrases that include them or noun phrases that appear in their immediate context. Candidate terms are further ranked using the following scoring function s, defined as:

$$s(\tau) = |\tau| \log f(\tau) + \alpha e_\tau \quad (1)$$

where τ is the candidate string, $|\tau|$ is the number of words of candidate τ, f is its frequency in the corpus, and e_τ is the number of terms that embed the candidate string τ. The parameter α is used to linearly combine the embeddedness score and is empirically set to 3.5. Only the best 20 expertise topics are stored for each document, where expertise topics are ranked based on their overall score $s(\tau)$ multiplied with their *tf-idf* score. To assign an expertise topic to authors, we compute the standard measure of relevance, *tf-idf*, for an aggregated document constructed by concatenating all their documents, which we call *tf-irf*. This approach follows the same overall process followed by the *LM1* approach. For the *expert finding task*, experts are ranked based on the topic score $s(\tau)$ multiplied with *tf-irf*. In the case of the *expert profiling task*, we implemented two approaches. The first approach (*DC*) ranks expertise topics for researchers based on the number of documents authored by them, that have the expertise topic stored as a keyphrase. The second approach (*TC*) considers the relevance score *tf-irf* as well, by multiplying the two scores.

[12]ACM Computing Classification System: http://www.acm.org/about/class/1998/

Table 3: Expert profiling results for the language modelling approach (LM) and the topic centric approach (TC)

Dataset	Measure	LM1	LM2	TC
CL	MAP	0.0256	0.0233	**0.0392**
	MRR	0.1857	0.2044	**0.2767**
	P@5	0.0637	0.0903	**0.1262**
SW	MAP	0.0082	0.0088	**0.0369**
	MRR	0.1271	0.1161	**0.3437**
	P@5	0.0482	0.0426	**0.1786**
IR	MAP	0.1052	**0.1679**	0.0879
	MRR	**0.3761**	0.3677	0.3364
	P@5	0.1831	**0.2197**	0.1737

5. BASELINE EXPERIMENTAL RESULTS

The considerable number of expertise topics annotated for the CL and SW datasets allows us to evaluate the precision of expertise topics ranked using the scoring function in Equation 1. The results can be seen in Figure 1 and Figure 2 respectively, where the candidate terms ranked using the topic-centric approach (*TC*) are compared with the same candidates ranked based on their *tf-idf* score. The *TC* approach outperforms the more simple *tf-idf* method, especially at the top of the ranked list on both datasets. The results for the expert finding baselines are presented in Table 2, in terms of mean average precision (*MAP*), mean reciprocal rank (*MRR*), and precision at top 5 (*P@5*). The *LM1* approach outperforms the *LM2* approach for the expert finding task, but both meth-

ods achieve much lower results than the topic-centric approaches *DC* and *TC*. Combining the *DC* baseline with the *tf-irf* relevance measure in the *TC* approach, brings only moderate improvements. The same performance measures are used for the expert profiling results shown in Table 3. The *TC* approach outperforms again the *LM1* and *LM2* approaches on the SW and CL datasets, but not on the IR dataset. The expert profiling task is easier for all the systems, when a smaller number of gold standard topics are available.

6. CONCLUSIONS

In this paper, we introduced three benchmark datasets for expert search that use information about workshops to automatically gather relevance assessments. These data sets cover the Information Retrieval, Computational Linguistics and Semantic Web domains, instead of single institutions, as done in previous datasets. In this way, a much larger number of topic-person associations can be identified, allowing a fine-grained evaluation of expert search. An additional contribution of this work is to present baseline results for a language modelling and a topic-centric approach. Our experiments lead to the conclusion that expertise topics can be extracted reasonably well using a term extraction approach. Also, topic-centric approaches generally outperform language modelling approaches when applied to a domain-specific collection and not to an entire organisation. Collecting relevance assessments for expert search from information about workshop organizers and PC members has its limitations. In some cases, committee members are selected based on their reputation alone, independent of their areas of interest and expertise. At the same time, workshops that have a broad scope, covering multiple research areas, will introduce noise. This is because all the PC members will be associated with each of these areas, although they might be experts only on a subset of topics. Our plans for future work include extending the IR dataset with a larger number of workshops. At the same time, we plan to consider additional baselines based on impact and co-authorship.

7. ACKNOWLEDGMENTS

We would like to thank Wei Lu for his help in creating an initial version of the IR test collection. This work has been funded in part by the European Union under Grant No. 258191 for the PROMISE project, as well as by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

8. REFERENCES

[1] P. Bailey, A. P. de Vries, N. Craswell, and I. Soboroff. Overview of the TREC 2007 Enterprise Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC)*, 2007.

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

[3] R. Berendsen, K. Balog, T. Bogers, A. Van den Bosch, and M. De Rijke. On the Assessment of Expertise Profiles. *Journal of the American Society for Information Science*, July 2013.

[4] H. Biswas and M. Hasan. Using Publications and Domain Knowledge to Build Research Profiles: An Application in Automatic Reviewer Assignment. In *Proceedings of the 2007 International Conference on Information and Communication Technology (ICICT'07)*, pages 82–86, 2007.

[5] G. Bordea, S. Kirrane, P. Buitelaar, and B. O. Pereira. Expertise Mining for Enterprise Content Management. In *LREC*, pages 3495–3498, 2012.

[6] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: Proceedings of the Twelfth International Conference on Information and Knowledge Management*, pages 528–531, New Orleans, LA, 2003.

[7] H. Deng, I. King, and M. R. Lyu. Formal Models for Expert Finding on DBLP Bibliography Data. In *ICDM '08: Proceedings of the Eighth IEEE International Conference on Data Mining*, pages 163–172. IEEE, 2008.

[8] S. T. Dumais and J. Nielsen. Automating the Assignment of Submitted Manuscripts to Reviewers. In *SIGIR '92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 233–244, New York, NY, USA, 1992. ACM.

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

[10] K. Hofmann, K. Balog, T. Bogers, and M. de Rijke. Contextual Factors for Finding Similar Experts. *Journal of the American Society for Information Science*, 61(5):994–1014, 2010.

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

[12] D. Mimno and A. McCallum. Expertise Modeling for Matching Papers with Reviewers. In *SIGKDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 500–509, 2007.

[13] M. A. Rodriguez and J. Bollen. An Algorithm to Determine Peer-Reviewers. In *'08: Proceedings of the Seventeenth International Conference on Information and Knowledge Management*, pages 319–328. ACM, 2008.

[14] E. Smirnova and K. Balog. A User-oriented Model for Expert Finding. In *Proceedings of the 33rd European conference on Advances in information retrieval*, ECIR'11, pages 580–592, Berlin, Heidelberg, 2011. Springer-Verlag.

[15] I. Soboroff, A. P. de Vries, and N. Craswell. Overview of the TREC 2006 Enterprise Track. In *In The Fifteenth Text Retrieval Conference (TREC 2006). NIST*, 2006.

[16] M. Stankovic, J. Jovanovic, and P. Laublet. Linked Data Metrics for Flexible Expert Search on the Open Web. In *Proceedings of the 8th extended semantic web conference on The semantic web: research and applications - Volume Part I*, ESWC'11, pages 108–123, Berlin, Heidelberg, 2011. Springer-Verlag.

[17] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. ArnetMiner: Extraction and Mining of Academic Social Networks. In *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 990–998, New York, NY, USA, 2008. ACM.

[18] D. Yarowsky and R. Florian. Taking the Load off the Conference Chairs: Towards a Digital Paper-Routing Assistant. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, pages 220–230, 1999.