Evaluation of Retrieval Algorithms for Expertise Search

Gaya K. Jayasinghe
Data61, CSIRO
Melbourne, VIC, Australia
gaya.jayasinghe@csiro.au

Sarvnaz Karimi
Data61, CSIRO
Marsfield, NSW, Australia
sarvnaz.karimi@csiro.au

Melanie Ayre
Data61, CSIRO
Melbourne, VIC, Australia
melanie.ayre@csiro.au

ABSTRACT

Evaluation of expertise search systems is a non-trivial task. While in a typical search engine the responses to user queries are documents, the search results for an expertise retrieval system are people. The relevancy scores indicate how knowledgeable they are on a given topic. Within an organisation, such a ranking of employees could potentially be difficult as well as controversial. We introduce an in-house capability search system built for an organisation with a diverse range of disciplines. We report on two attempts of evaluating six different ranking algorithms implemented for this system. Evaluating the system using relevance judgements produced in each of the two attempts leads to an understanding of how different methods of collecting judgements on people's expertise can lead to different effectiveness of algorithms.

Keywords

Evaluation, Expertise search, Capability discovery, People ranking.

1. INTRODUCTION

Expert finding is the process of finding human experts for a given topic of expertise. A search system that facilitates such a process is referred to as an expertise retrieval system. Expertise retrieval has been studied extensively in the past two decades in the information retrieval community. Recently, however, attention to this area is rising again. Talent acquisition and expert finding using an aggregation of public profiles, including those on social networks (e.g., [7, 14]) is of interest to recruitment agencies as well as internally in different companies. We introduce our system NETWORK [8] that is implemented within a large geographically dispersed company of over 5,000 employees. NETWORK is developed to satisfy the need for a system that allows its users to explore and analyse networks of employees and their expertise within various collaborative contexts. These contexts include organisational units such as teams and projects, their outputs such as products, publications or patents, as well as the use of facilities. NETWORK should provide expertise retrieval for a variety of disciplines and job descriptions, including managers, researchers, engineers, finance officers, human resources and support staff. For such a system, we focused on two aspects: (1) effective retrieval algorithms that work well on the data available in the organisation; and (2) evaluation of these algorithms.

We first report on the search methodologies implemented in NETWORK. Here, we only focus on the expertise retrieval service of the NETWORK system, ignoring other components such as visualisation and geographical maps. We also note that our problem is restricted to the topical aspect of expertise search and social and cognitive aspects are outside the scope of this paper. We then focus on two evaluation settings followed by evaluation challenges for this system. The majority of the published work in this area evaluates its proposed methods using either standard TREC settings through enterprise track datasets [1, 2, 6, 12] or one of the public datasets which are mostly focused on expertise retrieval in academia. While this is a valid approach, we still need to evaluate our in-house system using the data that it is built on and in the setting that it is used. To the best of our knowledge, the process of setting up evaluation of an in-house capability search system has not been the focus of any previous work. Our work explores the first steps to dive into this evaluation problem, reporting on two settings of self-assessment and assessment by colleagues.

2. RELATED WORK

Balog et al. [3] define expertise retrieval systems as consisting of two main components: *expert finding*, which helps to answer information needs such as "find me someone who is an expert on X" and *expert profiling* that helps to answer queries such as "in which topics is this person an expert". Our work is largely focused on expert finding, but our system provides detailed information and visualisations on each expert, including their fields of research, experience, qualifications, activities, biography, a network of colleagues, affiliations, current projects, and their location on a map.

2.1 Expert Ranking

In an earlier work, Serdyukov [11] divided expert finding methods into four categories: profile-based, document-based, window-based and graph-based methods. In profile-based methods, a personal profile is built for each candidate expert by merging all documents linked to the person and indexed for retrieval. Document- and window-based methods analyse text within more granular and less ambiguous levels such as documents or text windows. These methods assume that the individual relevance scores of documents or text windows to a topic add up to the expertise level of a person on the topic. Graph-based methods score candidate experts based on their centrality for a given topic in a network. For each category a number of different methods have been studied so far. Pioneer systems such as P@NOPTIC which was proposed in 2001 [5] were mostly profile-based. Document-based systems that use lan-

ADCS '16, December 05 - 07, 2016, Caulfield, VIC, Australia

ISBN 978-1-4503-4865-2/16/12. . . $15.00

DOI: http://dx.doi.org/10.1145/3015022.3015035

guage models and data fusion methods [9] have been introduced later. More recently a number of machine learning-based methods including learning to rank [10] have been proposed. In our system, we consider a number of settings including profile-based methods, and combinations of profile and graph-based methods.

2.2 Expert Finding Evaluation

Evaluation of expertise search systems has generally followed the TREC framework, starting from the TREC Enterprise track in 2005 which lasted for four years and included expert search as one of the tasks [1, 2, 6, 12]. Balog et al. [3] lists evaluation datasets and metrics popular in the published research in this area. They state that standard TREC metrics including Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), Normalised Discounted Cumulated Gain (NDCG), and Precision at rank N (P@N) are the main metrics used so far. They argue that in this area having relevant hits at the top of the ranked lists is very important, because following up on a false recommendation from an expertise retrieval system can be very expensive.

Apart from TREC Enterprise collections, of which at this date only those from 2005 and 2006 are publicly available, there are a number of other public datasets available such as ArnetMiner [13] and Yahoo! Answers. While these datasets are a valuable resource for evaluation of expertise search systems, there still exists a need for an in-house system to be evaluated on its own dataset. Literature, however, does not provide much guidance on how such in-house evaluation should be carried out. Our work focuses on this.

3. NETWORK SYSTEM

NETWORK [8] is a capability search system designed to help its users explore and analyse networks of people and their expertise within various collaborative contexts such as projects or the use of facilities (e.g., a Cytometry facility). NETWORK uses heterogeneous corporate datasets of peoples' profile pages; publication repositories; and internal organisational data to profile all employees. These employees have a variety of roles, such as managers, engineers, scientists, human resources, librarians, and financial officers, covering a broad range of experience and disciplines. NETWORK collates and links together information on the employees, creates representative profiles of experts in all areas, and provides the following services:

1. Ability for users to search and identify people with relevant expertise to a topic of interest;
2. Searching for people by name to browse their profiles;
3. Searching over business units for their capabilities;
4. Searching past or current projects linked to specific topics; and
5. Identifying experts on a topic that are located within a specific geographical region.

3.1 Datasets and Profiling

Unlike document retrieval, in an expertise retrieval system retrieved items are people. Therefore, candidates can be represented using a *profile* that is representative of their expertise. Evidence of expertise, however, can be collected from a variety of sources such as a company's Intranet or social media. Our profiling method is different to some of the past work in that a skills matrix [3] is created. This is to avoid creating a fixed set of areas of expertise, because of the broad diversity of employees in the organisation. To find evidence of knowledge area, we aggregate five different data types available to our system: employees, publications, organisations and organisational sub-units such as projects and teams, sites, and facilities. This data is transformed to an RDF dataset of approximately 1.1 million triples describing organisational data, publications and peoples' profiles, and contributions to software repositories.

Our system creates four different kinds of profiles: people, project, publication, and organisational unit. An overview of the information each of these profiles contains is shown below as vectors. We index this information using the Solr search engine. The scoring of expertise in each field is done in the ranking stage where different evidence of expertise is combined. NETWORK has indexed 24,983 people, 10,325 projects, 48,607 publications, and 1,202 organisational units.

Person = <ID, Organisation name, Position name, Title, First name, Last name, Site name, Street, Longitude, Latitude, State, Postcode, Telephone, Email, Organisation URI, Active, Personnel number, PageRank score, Publications, Publication URIs, Projects, Project URIs, Type>

Project = <ID, Profile, Unit name, Project name, WBS-Code, PageRank score, Leader URI, Leader title, Leader last name, Leader first name, Total expenditure, Publications, Publication URIs, Type, Active, Start date, End date, Client names>

Publication = <ID, Conference date, Publisher, Conference location, Classification, Title, Type, Abstract, Author names, Author URIs, Outcome, WBS-Code, Conference name, PageRank Score, Published date, Pages, Access, Source, Year of publication, Keywords, Journal title, Publication volume, Publication issue>

Organisational Unit = <ID, Unit name, Unit URI, Parent unit name, Parent unit URI, Type, Leader URI, Leader title, Leader first name, Leader last name, Persons, Persons URIs, PageRank score>

3.2 Ranking Experts

We implement six different information retrieval algorithms to rank experts for a given topic. The first class of algorithms only use textual information, and consist of two approaches, namely a BM25 Model and a BM25 model without document length normalisation (BM25 (No LN)). The second class of algorithms model the experts' network of relationships built by co-authoring publications, working in project teams, and being in organisational units in addition to textual information in profiles. Researchers have used similar strategies to explore academic social networks [13]. We implement two such strategies that utilise graph-based information. The first strategy employs PageRank. The second strategy consists of two steps. Firstly, it ranks all other profiles (publications, projects, and business units) except for expert profiles . In the second step, the relevance scores of ranked profiles are used to score experts. We call this algorithm Evidence-Based (EB). The Evidence-Based method is similar to the document-based methods for expertise retrieval, except that other types of profiles linked in a graph are analysed instead of documents related to people. Each strategy that uses experts' network of relationships is combined with two text retrieval strategies (BM25 and BM25 (No LN)).

4. EVALUATION

Evaluation of NETWORK, or any expertise retrieval system, is challenging for a number of reasons: (1) measuring expertise in a field is subjective; (2) different criteria can be applied to measure expertise within a multidisciplinary organisation, meaning that a single measure does not fit for all the staff; (3) ranking of people in comparison to each other for a specific field is not straightforward and sometimes can be contentious, for example one may argue that organisational ranks should be included while others do not agree; and (4) different quantities and different kinds of information is available for different employees in an organisation. We evaluated our system in two different settings: self-assessment and colleague assessment. Details are as below.

4.1 Self-Assessment

Following the evaluation setting proposed by Berendsen et al. [4], we assume that every employee knows their own expertise better than anyone else. We also assume that they are generally aware of colleagues with similar expertise. Therefore, we asked a group of volunteer employees to fill in a form that asked them to list their own areas of expertise and the corresponding list of current employees who they know that they share those expertise. We also asked them, if they can, to rank their lists based on level of expertise. Fourteen people participated. From them, we collected a list of 95 different expertise names which we refer to as queries. For each query, we had an average of 3 people listed. Queries contained between one and four words. Examples of these queries include: *climate adaptation*, *live software training*, *qualitative data analysis*, and *data mining*. Almost no one ranked their people list for any of the listed expertise, despite our assurance that the collected data will be kept anonymous. Although this method leads to relatively reliable expertise assigned to individuals, it lacks judgement on many others with the same expertise.

4.2 Colleague Assessment

In a second attempt to create a ground truth dataset for evaluating our system, we asked for volunteers within the organisation to participate in user study sessions to provide relevance judgements on the output of our ranking algorithms. We refer to these volunteers as *judges*. We used the topics that were suggested in the self-assessment study, with the assumption that they are partially representative of expertise within the organisation. We queried the indexed profiles using the six different algorithms implemented for NETWORK and added their top 10 answers to the pool for each given query. Duplicates were removed. The final set was randomly ordered. Using a web interface, judges were presented with a topic of expertise and a list of people with links to their profiles. They were told that they can also use other sources of information they can find on the web or internal company pages for their judgements. For example they could refer to their LinkedIn profiles. Three options were provided for the relevancy judgements: *Not relevant*, *relevant*, and *cannot decide*. There was no time limit set for the judges to finish their tasks and each judge could choose to provide judgements on one or more queries.

After advertising the task, twelve staff volunteered to participate, five of whom had participated in the self-assessment study. They provided judgements for 36 queries. Overall, they judged 1,356 hits, of which 220 were considered relevant, 808 irrelevant, and 328 undecided to their corresponding query. On average, each query had 6.1 relevant, 22.4 irrelevant, and 9.1 undecided judgements. There were also 328 items unjudged. From these 36 queries we only kept those for which the entire pool shown to the participants were judged, which left us with 20 queries with 144 relevant, 584 irrelevant and 187 undecided judgements.

Three major problems were identified in this user study:

1. Difficulty of judging experts;
2. Ambiguity of the topics; and
3. Unfamiliarity with the topics.

Judges noted that for some hits they could not reliably decide if a person is expert on the topic or not. No evidence pro or contra was found in their profile. Hence, we had 20.4% (187 out of 915) of the judgements as *cannot decide*. Also, some topics were considered ambiguous. That was when the judge had no comprehensive background on the given topic to reliably judge others on it, or the topic could have different meanings. For example, while one judge thought the topic *evaluation* is ambiguous and cannot be named as someone's expertise, the others disagreed. Another example was *PCL* as an expertise. The person who originally suggested this as their expertise was referring to *Point Cloud Library* which is software for image processing. However, PCL is an abbreviation for many more skills or areas of expertise. It could also bring people with expertise on *Posterior Cruciate Ligament* injury into the pool, and stand for various other words in different disciplines[1]. Therefore, the context in which one does the search on this system would make a difference in how the results are perceived. The final prominent reason for judges to decline to make judgements for a given topic was that they were not familiar with the topic and therefore decided that they cannot judge if someone is expert in it. For example, a topic *unity* (a cross-platform game engine) was not judged. Those topics that were too specific seemed to cause more trouble, whereas broader topics such as *project management* or *data mining* were easier to judge and got full judgement on their hits.

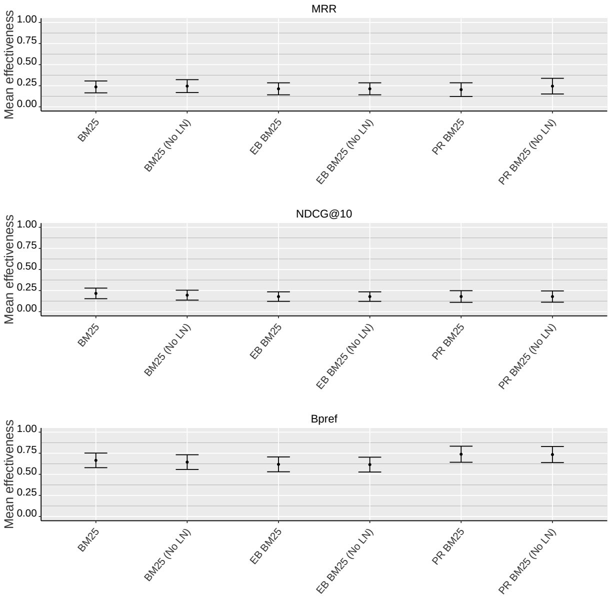

Figure 1: Effectiveness of ranking algorithms using self-assessments.

5. RESULTS AND DISCUSSION

We first evaluated our expertise ranking algorithms using the judgements that we obtained in the self-assessment exercise (see Section 4.1). The results are shown in Figure 1. The plots illustrate the mean and 95% confidence interval for evaluation metrics MRR, NDCG@10 and Bpref. The effectiveness of algorithms using MRR and NDCG@10 are poor for all algorithms. No clear wining strategy can be identified from this evaluation. However, a comparatively higher effectiveness is observed with Bpref. This shows that this evaluation suffers from lack of judgements for a large number of topics and experts. This makes the evaluation results less reliable, especially for those metrics that look at top n results when n is higher than 3 (average number of listed experts per query).

We then evaluated our ranking algorithms using queries and judgements from the colleague assessment. This time we had fewer queries but all hits in the top 10 results were judged. Therefore, using metrics such as NDCG@5 and NDCG@10 is more appropriate. Since in this system having high precision at higher ranks is

[1] https://en.wikipedia.org/wiki/PCL (Retrieved November 10, 2016)

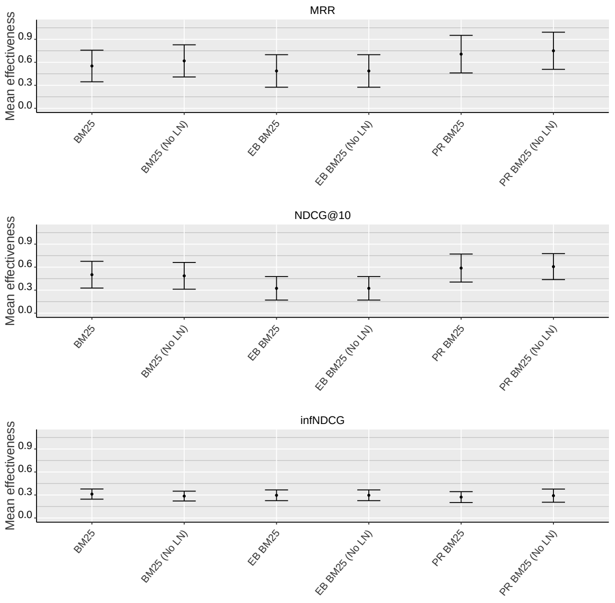

Figure 2: Effectiveness of ranking algorithms using judgements from the user study.

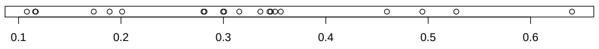

Figure 3: Variation in effectiveness (infNDCG) with BM25 ranking algorithm for 20 topics.

important, we also report MRR. Another metric that we can use for this particular setting is infNDCG [15]. Results are shown in Figure 2. As other evaluation metrics other than the infNDCG cannot handle a relevance level of *cannot decide*, we considered them as not relevant.

We depict the fluctuations in the results of BM25 algorithm using the infNDCG metric in Figure 3. Each circle represents a query. A wide variation in effectiveness for different queries is observed. For example, query *climate adaption*, performed worst with 0.1081 (infNDCG) and query *machine translation* performed best with 0.6404 (infNDCG).

6. CONCLUSIONS AND FUTURE WORK

NETWORK currently profiles an organisation with 5,000 employees working across a broad range of subject disciplines. It incorporates an expertise search engine as a core components. We implemented six different retrieval algorithms and evaluated them using two sets of relevance judgements obtained in two different settings: a self-assessment study and a user study to assess colleagues. Our observations were: (1) PageRank-based algorithms that use the internal organisation network show more promise than evidence-based methods that rank experts using different evidence prior to using their profile information; (2) effective design of evaluation of an in-house expertise retrieval system requires more research. While self-assessments are the most reliable means of knowing one's expertise, it is not feasible to get all the employees to provide such information. Judgements over colleagues' expertise is also difficult and leads to uncertainty in judgements.

Our system and its the evaluation are in their infancy, with a number of developments under way. NETWORK was released in October 2016 which enables us to now collect query logs and click data. Such data will be used to develop learning to rank and search result diversification algorithms as well as evaluating the system beyond collecting explicit relevance judgements. We are also considering the use of indirect information available to staff. Example of such implicit evidence are citations for researchers, participations in code repositories or online question answering for software engineers, and holding certificates such as a Payroll Practising Certificate for staff working in human resources administration.

Acknowledgement.

Authors would like to acknowledge the contributions of other project members, including David Ratcliffe, Alexander Krumpholz, Jonathan Yu, Hendra Wijaya, and Xavier Ho, in design and implementation of NETWORK. The project has Ethics Clearance (017/16 Network/Australia At Your Service) by CSIRO ethics committee.

References

[1] Peter Bailey, Nick Craswell, Arjen P. de Vries, and Ian Soboroff. Overview of the TREC 2007 enterprise track. In *TREC*, 2007.

[2] Krisztian Balog, Ian Soboroff, Paul Thomas, Peter Bailey, Nick Craswell, and Arjen P. de Vries. Overview of the TREC 2008 enterprise track. In *TREC*, 2008.

[3] Krisztian Balog, Yi Fang, Maarten de Rijke, Pavel Serdyukov, and Luo Si. Expertise retrieval. *FTIR*, 6(2–3):127–256, 2012.

[4] Richard Berendsen, Maarten de Rijke, Krisztian Balog, Toine Bogers, and Antal van den Bosch. On the assessment of expertise profiles. *JASIST*, 64(10):2024–2044, 2013.

[5] Nick Craswell, David Hawking, Anne-Marie Vercoustre, and Peter Wilkins. P@NOPTIC expert: Searching for experts not just for documents. In *Australasian World Wide web Conference*, pages 21–25, 2001.

[6] Nick Craswell, Arjen P. de Vries, and Ian Soboroff. Overview of the TREC-2005 enterprise track. In *TREC*, 2005.

[7] Ido Guy, Uri Avraham, David Carmel, Sigalit Ur, Michal Jacovi, and Inbal Ronen. Mining expertise and interests from social media. In *WWW*, pages 515–526, Rio de Janeiro, Brazil, 2013.

[8] Sarvnaz Karimi, Gaya Jayasinghe, David Ratcliffe, and Alexander Krumpholz. Network: A capability and people explorer system. In *CIKM Workshop on Data-Driven Talent Acquisition*, Indianapolis, IN, 2016.

[9] Craig Macdonald and Iadh Ounis. Voting for candidates: Adapting data fusion techniques for an expert search task. In *CIKM*, pages 387–396, Arlington, Virginia, USA, 2006.

[10] Catarina Moreira, Pável Calado, and Bruno Martins. Learning to rank academic experts in the dblp dataset. *Expert Systems: The Journal of Knowledge Engineering*, 32(4):477–493, 2015.

[11] Pavel Serdyukov. *Search for expertise : going beyond direct evidence*. PhD thesis, Enschede, June 2009. URL http://doc.utwente.nl/61651/.

[12] Ian Soboroff, Arjen P. de Vries, and Nick Craswell. Overview of the trec 2006 enterprise track. In *TREC*, 2006.

[13] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and mining of academic social networks. In *KDD*, pages 990–998, Las Vegas, NV, 2008.

[14] Kush R. Varshney, Vijil Chenthamarakshan, Scott W. Fancher, Jun Wang, Dongping Fang, and Aleksandra Mojsilović. Predicting employee expertise for talent management in the enterprise. In *KDD*, pages 1729–1738, New York, New York, USA, 2014.

[15] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. A simple and efficient sampling method for estimating AP and NDCG. In *SIGIR*, pages 603–610, Singapore, Singapore, 2008.