Finding the Right Supervisor: Expert-Finding in a University Domain

Fawaz Alarfaj, Udo Kruschwitz, David Hunter and Chris Fox
School of Computer Science and Electronic Engineering
University of Essex
Colchester, CO4 3SQ, UK
{falarf, udo, dkhunter, foxcj}@essex.ac.uk

Abstract

Effective knowledge management is a key factor in the development and success of any organisation. Many different methods have been devised to address this need. Applying these methods to identify the experts within an organisation has attracted a lot of attention. We look at one such problem that arises within universities on a daily basis but has attracted little attention in the literature, namely the problem of a searcher who is trying to identify a potential PhD supervisor, or, from the perspective of the university's research office, to allocate a PhD application to a suitable supervisor. We reduce this problem to identifying a ranked list of experts for a given query (representing a research area).

We report on experiments to find experts in a university domain using two different methods to extract a ranked list of candidates: a database-driven method and a data-driven method. The first one is based on a fixed list of experts (e.g. all members of academic staff) while the second method is based on automatic Named-Entity Recognition (NER). We use a graded weighting based on proximity between query and candidate name to rank the list of candidates. As a baseline, we use a system that ranks candidates simply based on frequency of occurrence within the top documents.

1 Introduction

The knowledge and expertise of individuals are significant resources for organisation. Managing this intangible resource effectively and efficiently constitutes an essential and very important task (Nonaka and Takeuchi, 1995; Law and Ngai, 2008). Approaching experts is the primary and most direct way of utilising their knowledge (Yang and Huh, 2008; Li et al., 2011). Therefore, it is important to have a means of locating the right experts within organisations. The expert-finding task can be categorised as an information retrieval task similar to a web search, but where the results are people rather than documents. An expert-finding system allows users to input a query, and it returns a ranked list of experts.

Here we look at a university context. We start with a real-world problem which is to identify a list of experts within an academic environment, e.g. a university intranet. The research reported here is based on an empirical study of a simple but effective method in which a system that applies the concept of expert-finding has been designed and implemented. The proposed system will contribute to provide an expert-search service to all of the university's stakeholders.

Expert-finding systems require two main resources in order to function: a list of candidates and a collection of data from which the evidence of expertise can be extracted. We present two approaches to address this problem, a database-driven and a data-driven method using NER. The main difference between the two methods is the way in which the candidates' list is constructed. In the database method, the candidates are simply selected from a known list of experts, e.g. the university's academic staff. In the NER method, the candidates are extracted *automatically* from the pages returned by an

Proceedings of the NAACL HLT 2012 Student Research Workshop, pages 1–6,
Montréal, Canada, June 3-8, 2012.

underlying search engine. This method promises to be more useful for finding experts from a wider (and possibly more up-to-date) range of candidates. Both methods apply the same ranking function(s), as will be discussed below.

This paper will survey related work in Section 2 and introduce the expert-finding task in a university domain in Section 3. The process of ranking experts will be discussed in Section 4. The evaluation will be described in Section 4, followed by a discussion of the experiment's results in Section 5.

2 Related Work

The expert-finding task addresses the problem of retrieving a ranked list of people who are knowledgeable about a given topic. This task has found its place in the commercial environment as early as the 1980's, as discussed in Maybury (2006); however, there was very limited academic research on finding and ranking experts until the introduction of the enterprise track at the 2005 Text REtrieval Conference (TREC) (Craswell et al., 2005).

When expert-finding we must know the experts' profiles, These profiles may be generated manually or automatically. Manually created profiles may be problematic. If, for example, experts enter their own information, they may exaggerate or downplay their expertise. In addition, any changes of expertise for any expert requires a manual update to the expert's profile. Thus incurring high maintenance costs. An example of manually generated profiles is the work of Dumais and Nielsen (1992). Although their system automatically assigns submitted manuscripts to reviewers, the profiles of the reviewers or experts are created manually.

The alternative is to generate the profiles automatically, for example by extracting relevant information from a document collection. The assumption is that individuals will tend to be expert in the topics of documents with which they are associated. Experts can be associated with the documents in which they are mentioned (Craswell et al., 2001) or with e-mails they have sent or received (Balog and de Rijke, 2006a; Campbell et al., 2003; Dom et al., 2003). They can also be associated with their home pages or CVs (Maybury et al., 2001), and with documents they have written (Maybury et al., 2001; Becerra-Fernandez, 2000). Finally, some researchers use search logs to associate experts with the web pages they have visited (Wang et al., 2002; Macdonald and White, 2009).

After associating candidate experts with one or more of the kinds of textual evidence mentioned above, the next step is to find and rank candidates based on a user query. Many methods have been proposed to perform this task. Craswell et al. (2001) create virtual documents for each candidate (or employee). These virtual documents are simply concatenated texts of all documents from the corpus associated with a particular candidate. Afterwards, the system indexes and processes queries for the employee's documents. The results would show a list of experts based on the ten best matching employee documents. Liu et al. (2005) have applied expert-search in the context of a community-based question-answering service. Based on a virtual document approach, their work applied three language models: the query likelihood model, the relevance model and the cluster-based language model. They concluded that combining language models can enhance the retrieval performance.

Two principal approaches recognised for expert-finding can be found in the literature. Both were first proposed by Balog et al. (2006b). The models are called the candidate model and the document model, or Model 1 and Model 2, respectively. Different names have been used for the two methods. Fang and Zhai (2007) refer to them as 'Candidate Generation Models and Topic Generation Models'. Petkova and Croft (2006) call them the 'Query-Dependent Model' and the 'Query-Independent Model'. The main difference between the models is that the candidate-based approaches (Model 1) build a textual representation of candidate experts, and then rank the candidates based on the given query, whereas the document-based approaches (Model 2) first find documents that are relevant to the query, and then locate the associated experts in these documents.

Balog et al. (2006b) have compared the two models and concluded that Model 2 outperforms Model 1 on all measures (for this reason, we will adopt Model 2).

As Model 2 proved to be more efficient, it formed the basis of many other expert-search systems (Fang

and Zhai, 2007; Petkova and Croft, 2007; Yao et al., 2008). Fang and Zhai developed a mixture model using proximity-based document representation. This model makes it possible to put different weights on different representations of a candidate expert (Fang and Zhai, 2007). Another mixture of personal and global language models was proposed by Serdyukov and Hiemstra (2008). They combined two criteria for personal expertise in the final ranking: the probability of generation of the query by the personal language model and a prior probability of candidate experts that expresses their level of activity in the important discussions on the query topic.

Zhu et al. (2010) claimed that earlier language models did not consider document features. They proposed an approach that incorporates: internal document structure; document URLs; page rank; anchor texts; and multiple levels of association between experts and topics.

All of the proposed frameworks assume that the more documents associated with a candidate that score highly with respect to a query, the more likely the candidate is to have relevant expertise for that query. Macdonald and Ounis (2008) developed a different approach, called the Voting Model. In their model, candidate experts are ranked first by considering a ranking of documents with respect to the users' query. Then, using the candidate profiles, votes from the ranked documents are converted into votes for candidates.

There have been attempts to tackle the expert-finding problem using social networks. This has mainly been investigated from two directions. The first direction uses graph-based measures on social networks to produce a ranking of experts (Campbell et al., 2003; Dom et al., 2003). The second direction assumes similarities among the neighbours in a social network and defines a smoothing procedure to rank experts (Karimzadehgan et al., 2009; Zhang et al., 2007).

Some have argued that it is not enough to find experts by looking only at the queries' without taking the users into consideration. They claim that there are several factors that may play a role in decisions concerning which experts to recommend. Some of these factors are the users' expertise level, social proximity and physical proximity (Borgatti and Cross, 2003; McDonald and Ackerman, 1998; Shami et al., 2008). McDonald and Ackerman (1998) emphasised the importance of the accessibility of the expert. They argued that people usually prefer to contact the experts who are physically or organisationally close to them. Moreover, Shami et al. (2008) found that people prefer to contact experts they know, even when they could potentially receive more information from other experts who are located outside their social network.

Woudstra and van den Hooff (2008) identified a number of factors in selecting experts that are related to quality and accessibility. They argued that the process of choosing which candidate expert to contact might differ depending on the specific situation.

Hofmann et al. (2010) showed that many of these factors can be modelled. They claimed that integrating them with retrieval models can improve retrieval performance. Smirnova and Balog (2011) provided a user-oriented model for expert-finding where they placed an emphasis on the social distance between the user and the expert. They considered a number of social graphs based on organisational hierarchy, geographical location and collaboration.

3 Expert-Finding in a University

In any higher educational institution, finding an appropriate supervisor is a critical task for research students, a task that can be very time consuming, especially if academics describe their work using terms that a student is not familiar with. A searcher may build up a picture of who is likely to have the relevant expertise by looking for university academic staff who have written numerous documents about the general topic, who have authored documents exactly related to the topic, or who list the topic as one of their research interest areas. Automating this process will not only help research students find the most suitable supervisors, but it also allow the university to allocate applications to supervisors, and help researchers find other people interested in the particular topics.

3.1 Method

The two approaches we apply, database-driven, and data-driven using NER[1] are illustrated in Figure 1.

[1] We use OpenNLP to identify named entities.

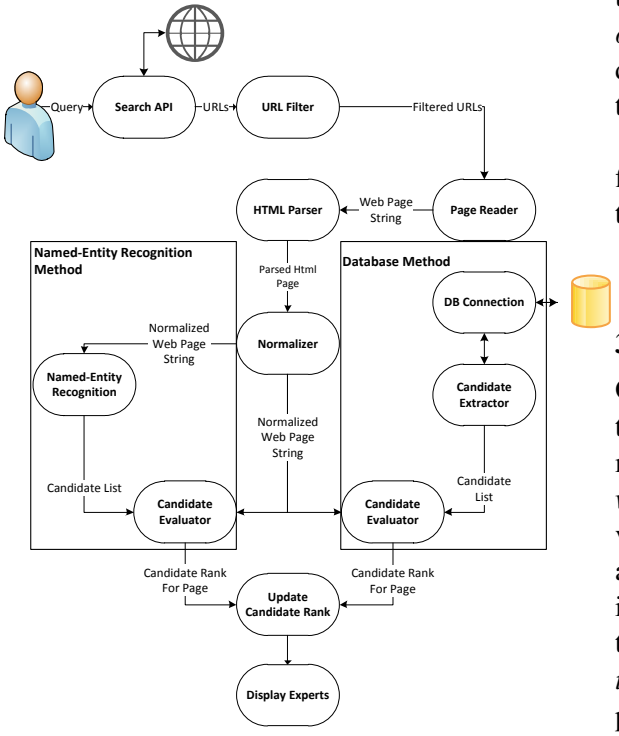

Figure 1: System Architecture.

The main difference between the two methods is the way in which the candidates' list is constructed. We argue that each method has its advantages. In the database method, the candidates are simply the university's academic staff. This avoids giving results unrelated to the university. It would be appropriate if the aim is to find the experts from among the university academics. In the data-driven method, the candidates are extracted from the pages returned by the underlying search engine. The experts found by this method are not necessarily university staff. They could be former academics, PhD students, visiting professors, or newly appointed staff.

Both methods apply the same ranking functions, one baseline function which is purely based on frequency and one which takes into account proximity of query terms with matches of potential candidates in the retrieved set of documents.

3.2 The Baseline Approach

The baseline we chose for ranking candidates is the frequency of appearance of names in the top twenty retrieved documents. The system counts how many times the candidate's name appears in the document $d(cc)$. Then it calculates the candidate metric cm by dividing the candidate count $d(cc)$ by the number of tokens in the document $d(nt)$.

Equation 1 defines the metric, where cm is the final candidate's metric for all documents and n is the number of documents.

$$cm = \sum_{d=1}^{n} \frac{d(cc)}{d(nt)} \tag{1}$$

3.3 Our Approach

Our approach takes into account the proximity between query terms and candidate names in the matching documents in the form of a *distance weight*. This measure will adds a *distance weight* value to the main candidate's metric that was generated earlier. Similar approaches have been proposed in the literature for different expert search applications Lu et al. (2006); Cao et al. (2005). The *distance weight* will be higher whenever the name appears closer to the query term, within a +/- 10 word window.

We experiment with two different formulae. The first formula is as follows:

$$cm1 = \sum_{i=1}^{n}\sum_{j=1}^{m}(cm + \frac{1}{\beta * \alpha_{ij}}), \alpha_{ij} = \begin{cases} d_{ij} & \text{if } d_{ij} \leq 10 \\ 0 & \text{Otherwise} \end{cases} \tag{2}$$

where n is the number of times the candidate's name has been found in the matching documents, m is the number of times the (full) query has been identified, and d_{ij} is the distance between the name position and query position (β has been set empirically to 3). The second formula is:

$$cm2 = \sum_{i=1}^{n}\sum_{j=1}^{m}(cm + \frac{1}{cm^{\alpha_{ij}}}), \alpha_{ij} = \begin{cases} d_{ij} & \text{if } d_{ij} \leq 10 \\ 0 & \text{Otherwise} \end{cases} \tag{3}$$

This equation is designed to return a smaller value as the distance x increases, and to give the candidate with lower frequency a higher weight.

In both cases, candidates are ranked according to the final score and displayed in order so that the candidates who are most likely to be experts are displayed at the top of the list.

4 Evaluation

As with any IR system, evaluation can be difficult. In the given context one might argue that precision

is more important than recall. In any case, recall can be difficult to measure precisely. To address these issues we approximate a *gold standard* as follows. We selected one school within the university for which a page of research topics with corresponding academics exists In this experiment we take this mapping as a complete set of correct matches. In this page, there are 371 topics (i.e. potential queries) divided among 28 more general research topics. Each topic/query is associated with one or more of the school's academic staff. It is presumed that those names belong to experts on the corresponding topics.

Table 1 illustrates some general topics with the number of (sub)topic they contain. Table 2 list some of the topics.

Topic	**N**
Analogue and Digital Systems Architectures	2
Artificial Intelligence	26
Audio	12
Brain Computer Interface	18
Computational Finance Economics and Management	1
Computational Intelligence	10
...	...

Table 1: Distribution of topics - **N** denotes the number of topics for the corresponding general topic area.

High-Speed Lasers And Photodetectors
Human Behaviour And The Psychology
Human Motion Tracking
Human-Centred Robotics
Hybrid Heuristics
Hybrid Intelligent Systems Which Include Neuro-Fuzzy Systems
Hypercomplex Algebras And Fourier Transforms
Hypercomplex Fourier Transforms And Filters

Table 2: Some topics/queries

The measure used to test the system is *recall* at the following values {3, 5, 7, 10, 15, 20}. We also measure Mean Average Precision at rank 20 (MAP@20).

5 Results and Discussion

Table 3 shows the system results where BL is the baseline result. There are two main findings. First of all, the database-driven approach outperforms the data-driven approach. Secondly, our approach which applies a grading of results based on proximity between queries and potential expert names significantly outperforms the baseline approach that only considers frequency, that is true for both formulae we apply when ranking the results (using paired t-tests applied to MAP with p<0.0001). However, the differences between $cm1$ and $cm2$ tend not to be significantly different.

	BL		$cm1$		$cm2$	
	NER	DB	NER	DB	NER	DB
R@3	0.47	0.48	0.49	0.76	0.58	0.79
R@5	0.56	0.60	0.58	0.83	0.68	0.86
R@7	0.61	0.64	0.62	0.87	0.72	0.88
R@10	0.65	0.69	0.68	0.89	0.78	0.90
R@15	0.69	0.72	0.74	0.91	0.80	0.91
R@20	0.71	0.75	0.76	0.92	0.82	0.93
MAP	0.20	0.28	0.50	0.61	0.52	0.66

Table 3: Performance Measures

It is perhaps important to mention that our data is fairly clean. More noise would make the creation of relational database more difficult. In that case the data-driven approach may become more appropriate.

6 Conclusion

The main objective of this work was to explore expert-finding in a university domain, an area that has to the best of our knowledge so far attracted little attention in the literature. The main finding is that a database-driven approach (utilising a fixed set of known experts) outperforms a data-driven approach which is based on automatic named-entity recognition. Furthermore, exploiting proximity between query and candidate outperforms a straight frequency measure.

There are a number of directions for future work. For example, modelling the user background and interests could increase the system's effectiveness. Some more realistic end-user studies could be used to evaluate the systems. Consideration could be given to term dependence and positional models as in Metzler and Croft (2005), which might improve our proximity-based scoring function. Finaly, our gold standard collection penalises a data-driven approach, which might offer a broader range of experts. We will continue this line of work using both technical evaluation measures as well as user-focused evaluations.

References

K. Balog and M. de Rijke. Finding experts and their details in e-mail corpora. In *Proceedings of the 15th international conference on World Wide Web*, pages 1035–1036. ACM, 2006a.

K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM, 2006b.

I. Becerra-Fernandez. Facilitating the online search of experts at NASA using expert seeker people-finder. In *Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management (PAKM)*, Basel, Switzerland, 2000.

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

C.S. Campbell, P.P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531. ACM, 2003.

Y. Cao, J. Liu, S. Bao, and H. Li. Research on expert search at enterprise track of trec 2005. In *14th Text Retrieval Conference (TREC 2005)*, 2005.

N. Craswell, D. Hawking, A.M. Vercoustre, and P. Wilkins. P@ noptic expert: Searching for experts not just for documents. In *Ausweb Poster Proceedings, Queensland, Australia*, 2001.

N. Craswell, A.P. de Vries, and I. Soboroff. Overview of the TREC-2005 enterprise track. In *TREC 2005 Conference Notebook*, pages 199–205, 2005.

B. Dom, I. Eiron, A. Cozzi, and Y. Zhang. Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 42–48. ACM, 2003.

S.T. Dumais and J. Nielsen. Automating the assignment of submitted manuscripts to reviewers. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 233–244. ACM, 1992.

H. Fang and C.X. Zhai. Probabilistic models for expert finding. *Advances in Information Retrieval*, pages 418–430, 2007.

K. Hofmann, K. Balog, T. Bogers, and M. de Rijke. Contextual factors for finding similar experts. *Journal of the American Society for Information Science and Technology*, 61(5):994–1014, 2010.

M. Karimzadehgan, R. White, and M. Richardson. Enhancing expert finding using organizational hierarchies. *Advances in Information Retrieval*, pages 177–188, 2009.

C. Law and E. Ngai. An empirical study of the effects of knowledge sharing and learning behaviors on firm performance. *Expert Systems with Applications*, 34(4):2342–2349, 2008.

M. Li, L. Liu, and C. Li. An approach to expert recommendation based on fuzzy linguistic method and fuzzy text classification in knowledge management systems. *Expert Systems with Applications*, 38 (7):8586–8596, 2011.

X. Liu, W.B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 315–316. ACM, 2005.

W. Lu, S. Robertson, A. MacFarlane, and H. Zhao. Window-based enterprise expert search. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*. NIST, 2006.

C. Macdonald and I. Ounis. Voting techniques for expert search. *Knowledge and information systems*, 16(3):259–280, 2008.

C. Macdonald and R.W. White. Usefulness of click-through data in expert search. In *SIGIR*, volume 9, pages 816–817, 2009.

M. Maybury, R. D'Amore, and D. House. Expert finding for collaborative virtual environments. *Communications of the ACM*, 44(12): 55–56, 2001.

M.T. Maybury. Expert finding systems. *MITRE Center for Integrated Intelligence Systems Bedford, Massachusetts, USA*, 2006.

D.W. McDonald and M.S. Ackerman. Just talk to me: a field study of expertise location. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 315–324. ACM, 1998.

D. Metzler and W.B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 472–479. ACM, 2005.

I. Nonaka and H. Takeuchi. *The knowledge creating company: How Japanese The knowledge creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press, New York, 1995.

D. Petkova and W.B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on*, pages 599–608. IEEE, 2006.

D. Petkova and W.B. Croft. Proximity-based document representation for named entity retrieval. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 731–740. ACM, 2007.

P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. *Advances in Information Retrieval*, pages 309–320, 2008.

N.S. Shami, K. Ehrlich, and D.R. Millen. Pick me: link selection in expertise search results. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1089–1092. ACM, 2008.

E. Smirnova and K. Balog. A user-oriented model for expert finding. *Advances in Information Retrieval*, pages 580–592, 2011.

J. Wang, Z. Chen, L. Tao, W.Y. Ma, and L. Wenyin. Ranking user's relevance to a topic through link analysis on web logs. In *Proceedings of the 4th international workshop on Web information and data management*, pages 49–54. ACM, 2002.

L. Woudstra and B. van den Hooff. Inside the source selection process: Selection criteria for human information sources. *Information Processing & Management*, 44(3):1267–1278, 2008.

K. Yang and S. Huh. Automatic expert identification using a text automatic expert identification using a text categorization technique in knowledge management systems. *Expert Systems with Expert Systems with Applications*, 34(2):1445–1455, 2008.

J. Yao, J. Xu, and J. Niu. Using role determination and expert mining in the enterprise environment. In *Proceedings of the 2008 Text REtrieval Conference (TREC 2008)*, 2008.

J. Zhang, J. Tang, and J. Li. Expert finding in a social network. *Advances in Databases: Concepts, Systems and Applications*, pages 1066–1069, 2007.

J. Zhu, X. Huang, D. Song, and S. Rüger. Integrating multiple document features in language models for expert finding. *Knowledge and Information Systems*, 23(1):29–54, 2010.