

# Finding Similar Experts

Krisztian Balog  
ISLA, University of Amsterdam  
Kruislaan 403, 1098 SJ Amsterdam  
kbalog@science.uva.nl

Maarten de Rijke  
ISLA, University of Amsterdam  
Kruislaan 403, 1098 SJ Amsterdam  
mdr@science.uva.nl

## ABSTRACT

The task of finding people who are experts on a topic has recently received increased attention. We introduce a different expert finding task for which a small number of example experts is given (instead of a natural language query), and the system's task is to return *similar experts*. We define, compare, and evaluate a number of ways of representing experts, and investigate how the size of the initial example set affects performance. We show that more fine-grained representations of candidates result in higher performance, and larger sample sets as input lead to improved precision.

## Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.4 [Information Systems Applications]: H.4.2 Types of Systems; H.4.m Miscellaneous

## General Terms

Algorithms, Measurement, Performance, Experimentation

## Keywords

Expert finding, Similar experts, Expert representation

## 1. INTRODUCTION

Increasingly, commercial systems and the information retrieval community pay attention to retrieving entities and not just documents. Web search engines offer facilities for searching specific types of entities, such as books, CDs, restaurants. In 2005 and 2006 the TREC Enterprise track provided another example with the *expert finding* task, where systems return a ranked list of person names in response to a query. Here, people are being sought that are knowledgeable about a given topic, described in natural language.

We address a different expert finding task: we do not assume that the person seeking for experts supplies an explicit description of the area in which she seeks expertise (she might simply not be sufficiently knowledgeable). Instead, our user provides a small number of example experts—people that she knows personally or by reputation—, and the system has to return *similar experts*.

Finding similar experts (or, more generally, similar people), differs from finding similar documents in a number of ways. Most importantly, experts are not represented directly (as retrievable units such as documents), and we need to identify them indirectly through occurrences in documents. This gives rise to our main research

question: what are effective ways of representing candidate experts for our finding similar experts task? We define, compare, and evaluate four ways of representing experts: through their collaborations, through the documents they are associated with, and through the terms they are associated with (as a set of discriminative terms or vector of weighted terms). Our second research question concerns the number of example experts provided by the user: how does the size of the sample set affect end-to-end performance?

An additional contribution consists of a method for generating example sets and the corresponding “complete” sets, against which our results are evaluated. We use the TREC 2006 expert finding queries as evidence of a person's expertise, and use this evidence to create sets of people that are all experts on the same topics.

In Section 2 we point to related work; in Sections 3 and 4 we describe the expert representations and notions of similarity used. Results are presented in Section 5 and conclusions in Section 6.

## 2. RELATED WORK

As a research area, automatic approaches to expert finding received a big boost from the introduction of the expert finding task at TREC 2005 [6]. Many approaches fall into one of two types: create a textual representation of the experts and estimate how probable a given expertise area is, or create a textual representation of the given expertise area and estimate how probable a candidate expert is. Various variations on the expert finding task have since been proposed, including *expert profiling*—given a person, return a ranked list of topics in which she is an expert [1].

The finding similar experts task that we address may be viewed as a list completion task. List queries are common types of web queries [5]. Their importance has been recognized by the TREC Question Answering track [7] (where systems return two or more instances of the class of entities that match a description) and by commercial parties (e.g., Google Sets allows users to retrieve entities that resemble the examples provided [4]). Ghahramani and Heller [3] developed an algorithm for completing a list based on examples using Bayesian inference techniques. A proposed list completion task will likely be run at INEX 2007 [2].

## 3. REPRESENTING CANDIDATES

We introduce four ways of representing a candidate *ca*:

**(WG)** As a set of people that *ca* is working with. We use organizational information, and  $WG(ca)$  is a set of working groups that *ca* is a member of.

**(DOC)** As a set of documents associated with *ca*.  $DOC(ca)$  denotes a set of documents in which *ca* appears (i.e. it contains *ca*'s name or e-mail address).

**(TERM)** As a set of terms extracted from  $DOC(ca)$ .  $TERM(ca)$  contains only the top discriminative terms (with highest TF-IDF value) for each document.

Sample set size	WG				DOC				TERM				TERMVECT			
	MRR	P@5	P@10	P@15	MRR	P@5	P@10	P@15	MRR	P@5	P@10	P@15	MRR	P@5	P@10	P@15
1	0.267	0.134	0.138	0.135	0.478	0.317	0.300	0.288	0.475	0.322	0.324	0.316	0.596	0.407	0.396	0.383
2	0.319	0.159	0.167	0.166	0.541	0.381	0.373	0.368	0.531	0.374	0.368	0.370	0.723	0.549	0.523	0.495
3	0.346	0.173	0.186	0.187	0.608	0.451	0.446	0.438	0.609	0.456	0.460	0.456	0.765	0.608	0.586	0.560
4	0.361	0.186	0.194	0.201	0.638	0.483	0.476	0.469	0.676	0.502	0.507	0.506	0.824	0.681	0.656	0.615
5	0.382	0.196	0.194	0.196	0.642	0.492	0.495	0.493	0.701	0.547	0.548	0.546	0.853	0.703	0.676	0.638

**Table 1: Results on the finding similar experts task, averaged over 426 sample sets (sample set size 1) or 1,000 samples (other sizes).**

(TERMVECT) As a vector of term frequencies, extracted from  $DOC(ca)$ . Terms are weighted using the TF-IDF value.

#### 4. MEASURING SIMILARITY

Let  $S = \langle ca_1, \dots, ca_n \rangle$  be the sequence of examples provided by the user. Given  $S$ , the score of candidate  $ca$  is computed using

$$score(ca) = \sum_{ca' \in S} sim(ca, ca'),$$

where  $sim(ca, ca')$  reflects the degree of similarity between candidates  $ca$  and  $ca'$ . The  $m$  candidates with the highest score are returned as output. Using the representations described above we compute similarity scores as follows. For the set-based representations (WG, DOC, TERM) we compute the Jaccard coefficient. E.g., similarity based on the (DOC) representation boils down to

$$sim(ca, ca') = \frac{|DOC_{ca} \cap DOC_{ca'}|}{|DOC_{ca} \cup DOC_{ca'}|}.$$

Similarity between vectors of term frequencies (TERMVECT) is estimated using the cosine distance:

$$sim(ca, ca') = \cos(\vec{t}(ca), \vec{t}(ca')) = \frac{\vec{t}(ca) \cdot \vec{t}(ca')}{\|\vec{t}(ca)\| \|\vec{t}(ca')\|},$$

where  $\vec{t}(ca)$  and  $\vec{t}(ca')$  denote the term frequency vectors representing candidate  $ca$  and  $ca'$ , respectively.

#### 5. EXPERIMENTAL EVALUATION

##### Experimental design.

For evaluation we use the TREC Enterprise test collections. The document collection is the W3C corpus (appr. 330.000 documents, 5.7GB). Names and e-mail addresses of 1092 expert candidates (W3C members) are given. Working group membership information (needed for  $WG$ ) is provided by the TREC 2005 expert finding topics and qrels.

To simulate the user’s input (a set of example experts) and to generate the corresponding “complete” set of similar experts that can be used as ground truth, we used the following algorithm. The algorithm generates random sets of experts, with size  $\geq n + m$ , where  $n$  is the size of the example set, and  $m$  is the minimal number of additional experts that belong to the same set. We write  $expert(ca, t)$  to denote that  $ca$  is an expert on topic  $t$ ; the TREC 2006 topics and qrels are used to define  $expert(ca, t)$ .

1. Select  $n$  candidates at random (the sample set  $S$ ), and put  $T = \{t \mid \forall ca \in S, expert(ca, t)\}$ . Repeat until  $T \neq \emptyset$ .
2.  $CA$  is the set of additional candidates who are experts on  $T$ :  $CA = \{ca \mid ca \notin S, \forall t \in T, expert(ca, t)\}$
3. The sample set  $S$  is valid, if  $|CA| \geq m$

We conducted experiments for various input sizes ( $n = 1, \dots, 5$ ). Our system is expected to complete the list with 15 additional candidates ( $m = 15$ ). For each  $n$ , we generated 1,000 input sets (except for  $n = 1$ , where the number of valid sets is only 426).

We measured the mean reciprocal rank of the first retrieved result (MRR), as well as precision at 5, 10, and 15.

##### Results.

In Table 1 we report on the results of our experiments. We have two important findings. First, more fine-grained representations of candidates consequently result in higher performance (for all measures). Second, concerning the size of the example set, we conclude that larger input samples lead to higher scores (for all measures). Our best representation (TERMVECT) delivers excellent performance, achieving MRR=0.853, P@5=0.703 (for  $n = 5$ ).

Interestingly, the (DOC) representation is similar in performance to (TERM) for very small example sets, but loses out on larger sets. Also, for (WG), (DOC), (TERM) the P@5, P@10, P@15 scores tend to be very similar, while for (TERMVECT) we clearly have  $P@5 > P@10 > P@15$ .

#### 6. CONCLUSIONS

We introduced an expert finding task for which a small number of example experts is given, and the system’s task is to return *similar experts*. We defined, compared, and evaluated four ways of representing experts: through their collaborations, through the documents they are associated with, and through the terms they are associated with (either as a set of discriminative terms or as a vector of term weights). Moreover, we introduced a method that generates and validates random example sets, and determines the “complete” set, against which our results are evaluated. We found that more fine-grained representations of candidates result in higher performance; a vector of weighted term frequencies, extracted from the documents associated with the person, is proven to be the most effective way of representing candidate experts. Finally, larger sample sets as input lead to better overall performance.

#### 7. ACKNOWLEDGEMENTS

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project number 220-80-001.

#### 8. REFERENCES

- [1] K. Balog and M. de Rijke. Determining expert profiles (with an application to expert finding). In *Proc. IJCAI '07*, pages 2657–2662, 2007.
- [2] A. de Vries and N. Craswell. XML entity ranking track, 2006. URL: <http://inex.is.informatik.uni-duisburg.de/2006/xmlSearch.html>.
- [3] Z. Ghahramani and K. A. Heller. Bayesian sets. In *NIPS 2005*, 2005.
- [4] Google, 2006. GoogleSets. URL: <http://labs.google.com/sets>, accessed on 04-10-2006.
- [5] D. E. Rose and D. Levinson. Understanding user goals in web search. In *WWW '04*, pages 13–19. ACM Press, 2004.
- [6] TREC. Enterprise track, 2005. URL: <http://www.ins.cwi.nl/projects/trec-ent/wiki/>.
- [7] E. Voorhees. Overview of the TREC 2004 question answering track. In *Proceedings of TREC 2004*, 2005. NIST Special Publication: SP 500-261.