Modeling Expert Finding as an Absorbing Random Walk

Pavel Serdyukov, Henning Rode, Djoerd Hiemstra
Database Group, University of Twente
PO Box 217, 7500 AE
Enschede, The Netherlands
{serdyukovpv, rodeh, hiemstra}@cs.utwente.nl

ABSTRACT

We introduce a novel approach to expert finding based on multi-step relevance propagation from documents to related candidates. Relevance propagation is modeled with an absorbing random walk. The evaluation on the two official Enterprise TREC data sets demonstrates the advantage of our method over the state-of-the-art method based on one-step propagation.

Categories and Subject Descriptors:
H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval.

General Terms:
Algorithms, Measurement, Performance, Experimentation.

Keywords:
Enterprise search, expert finding, random walks.

1. INTRODUCTION

Expert finding is a new rapidly growing Information Retrieval research area and a popular application domain [3]. An expert finding system ranks people within an organization by their ability to share the knowledge described by a short user query [6]. The analysis of personal expertise is usually based on measuring the degree of co-occurrence of personal identifiers with query terms in the organizational documents. The most popular and successful methods following this principle consider the weighted sum of retrieval scores of all documents related to a candidate as a measure of candidate's expertise [1, 5]. In other words, they rely on *one-step relevance probability propagation* from documents to candidates. The presented method, which is the main contribution of this paper, allows *multi-step relevance probability propagation* from documents even to those candidates that are not explicitly mentioned in these documents. The propagation is modeled with an absorbing random walk.

2. EXPERT FINDING AS AN ABSORBING RANDOM WALK

One of the most theoretically sound and effective methods, proposed by Balog et al. [1], considers the expertise degree of the candidate expert e to be calculated as:

$$Expertise(e) = \sum_{D \in Top} P(Q|D)P(e|D) \quad (1)$$

$$P(e|D) = \frac{a(e,D)}{\sum_{e'} a(e',D)}, \quad (2)$$

where $P(Q|D)$ is the probability that the document D generates the query Q, measuring the document relevance according to the probabilistic language modeling principle of IR [4], $P(e|D)$ is the probability of association between the candidate and the document, $a(e,D)$ is the non-normalized association score between the candidate and the document proportional to their strength of relation.

The inspiration for our method comes from the analysis of the above formulations. If we look at Formulas 1 and 2 we may notice that they describe a probabilistic process of manually finding an expert. Considering that the candidate e is an expert, they calculate the probability that a user finds it by making the following steps. The user selects a document among the ones appearing in the initial ranking, looks through the document, enlists all candidate experts mentioned in it and refers with the current information need to one of them. The probability of selecting a document equals its probabilistic relevance score since the user will most probably search for useful information and contacts of knowledgeable people in one of the top documents recommended by a search engine.

While obviously modeling the manual search for expertise, the described method is based on the strict unrealistic assumption that the user stops after the first step of moving from documents to candidate experts. However, a real-world user may proceed further by asking immediately found candidates to recommend other documents describing the topic and by finding new candidates in these documents. So, if we consider that the user started the walk at some document, it is usually possible to find even candidates not directly mentioned in this document.

In our approach we represent the search for an expert as *an absorbing random walk* in a document-candidate graph. We calculate the probability of finding a candidate if consider that this candidate is the required expert. The candidate node which we want to evaluate is only self-transient, since we assume it to be the final destination of the walk. This means that in contrast to the one-step approach, we calculate the probability of finding a certain candidate expert by making infinite (or any sufficient) number of steps in the graph. Formally speaking, we remove all outgoing edges from the measured candidate, add the self-transition edge

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

to it and use the following formulas iteratively:

$$P_0(D) = P(Q|D), P_0(e) = 0, \quad (3)$$

$$P_i(D) = \sum_{e \to D} P(D|e) P_{i-1}(e), \quad (4)$$

$$P_i(e) = \sum_{D \to e} P(e|D) P_{i-1}(D) + P_{i-1}(e) P^{self}(e|e) \quad (5)$$

Since, we also consider the probability of moving from a candidate to a document, we have $P(D|e) = \frac{a(e,D)}{\sum_{D'} a(e,D')}$. Finally, we consider that $Expertise(e)$ is proportional to the probability $P_\infty(e)$.

It should be mentioned that in strongly connected graphs with an absorbing node the probability of absorption in the infinity is effectively close to 1. However, the graphs we deal with are nearly uncoupled and, de facto, the probability of absorption for certain node depends only on the connectivity of the region it belongs to and on the total probability of relevance of neighboring document nodes.

Making the full run of iterations for each candidate is unnecessary. If we rewrite the above formulas in a matrix form, we get $\mathbf{p} = \mathbf{p}_0 \mathbf{A}^i$, where $\mathbf{A}$ consists of one-step transition probabilities and $\mathbf{A}^i$ contains probabilities of transitioning from one node to another in i steps. In our calculations we use matrix $\mathbf{B}$ containing probabilities of transitioning from each node to another *in the minimum number of steps*. We get this matrix by filling it with those elements from $\mathbf{A}^i$, which become non-zero after some next iteration. When no new element in $\mathbf{A}^i$ becomes non-zero after some iteration, the filling of $\mathbf{B}$ is finished. The probabilities used for candidate ranking are calculated as $\mathbf{p} = \mathbf{p}_0 \mathbf{B}$. Our method can be also regarded as a generalization of Formula 1, considering that $P(e|D)$ is the probability to transfer from the document D to the candidate e not in one step, but in the minimum sufficient number of steps.

3. EXPERIMENTS

We experiment with two data sets used by the TREC community in 2005-2007. The **W3C** collection represents the internal documentation of the World Wide Web Consortium (W3C). In our experiments we use the largest (1.85 GB), the most clean and structured part of the corpus, containing email discussions within the W3C. We experiment with 49 queries, the list of 1092 candidate experts and respective relevance (expertise) judgments used for TREC evaluations in 2006. The **CSIRO** collection is a crawl from Australia's national science agency's (CSIRO) web site. 50 queries with judgments made by retired CSIRO employees and 3500 candidates found in the collection were used for the evaluation. At the collection preparation stage, we extract associations between candidate experts and documents by searching for candidates email addresses and full names in the text of documents. For the CSIRO documents the association scores $a(e, D)$ are set uniformly to 1.0. The association scores for different email fields in case of W3C data are set as recommended by recent studies [2]. We had to retrieve 1500 documents from the W3C collection and just 50 documents from the CSIRO collection for the maximum performance of the baseline method.

We evaluate three methods: our method, the baseline method, described by Formulas 1 and 2, and the simplest of known methods, called Votes [5], which ranks candidates just by the number of top documents where they appear. All methods are compared using popular IR performance measures also used in official TREC evaluations: Mean Average Precision (MAP), precision at top 5 ranked candidate experts (P@5) and Mean Reciprocal Rank (MRR).

	W3C, 2006			CSIRO, 2007		
	MAP	MRR	P@5	MAP	MRR	P@5
Baseline	0.379	0.787	0.624	0.361	0.508	0.220
Votes	0.336	0.700	0.571	0.321	0.449	0.212
Our method	0.398	0.804	0.641	0.376	0.518	0.232

Table 1: Performance for all measures, both data sets and all tested methods

The performance of all tested methods for all measures is presented in Table 1. We see that our method outperforms the baseline method for all measures on both datasets. Considering the advantage of the baseline over the simplest Votes method and that the baseline is one of the most effective methods known, we may conclude that the improvements made by our method show the importance and potential of multi-step relevance probability propagation principle for expert finding. It was also clear from the experiments that relevance of documents situated farther than in 3 nodes from a measured candidate has almost no influence on its probability of absorption, since the probability of following such paths is too low.

4. CONCLUSIONS AND FUTURE WORK

We presented an expert finding method based on an absorbing random walk in a document-candidate graph. We showed that our method is a generalization of the popular one-step relevance propagation based method. It allows to propagate relevance appearing from retrieved documents even to not directly mentioned candidates. Experiments on TREC data originated from two large organizations demonstrated the advantage of the multi-step relevance propagation over the baseline one-step propagation.

In future we are going to evaluate the presented approach considering that the user is able to follow not only mutual links between documents and candidates, but also hyperlinks among documents and organizational links among candidate experts. It would be also interesting to study other, maybe non-probabilistic relevance propagation methods, to allow for weighting the influence of document neighborhoods of different levels on the degree of candidate's expertness.

5. REFERENCES

[1] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *SIGIR'07*, pages 551–558, 2007.

[2] K. Balog and M. de Rijke. Finding experts and their details in e-mail corpora. In *WWW'06*, 2006.

[3] N. Craswell, A. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *Proceedings of TREC'05*, 2005.

[4] D. Hiemstra. *Using Language Models for Information Retrieval*. Phd thesis, University of Twente, 2001.

[5] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM'06*, 2006.

[6] M. T. Maybury. Expert finding systems. Technical Report MTR06B000040, MITRE Corporation, 2006.