Exploiting Sequential Dependencies for Expert Finding

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 propose an expert finding method based on assumption of sequential dependence between a candidate expert and the query terms in the scope of a document. We assume that the strength of relation of a candidate to the document's content depends on its position in this document with respect to the positions of the query terms. The experiments on the official Enterprise TREC data demonstrate the advantage of our method over the method based on independence of query terms and persons in a document.

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, dependence modeling.

1. INTRODUCTION

The search for people with relevant expertise rather than for relevant documents is becoming more and more of a necessary everyday activity in organizations. Recently, the expert finding task attracted the close attention of IR research community and stays a part of the Enterprise track of the Text REtrieval Conference (TREC) [2] already since 2005. The majority of approaches proposed so far is based on inferring personal expertness by analyzing the degree of co-occurrence of personal identifiers with the query terms in top retrieved organizational documents. The state-of-the-art methods relying on this principle in fact consider that the candidate's expertness can be measured by the sum of retrieval scores of related documents [1, 5]. In this paper we propose to take not only the fact of co-occurrence into account, but also the important property of this co-occurrence: the sequential order of a candidate's identifier and the query terms mentioned in a document.

2. EXPLOITING SEQUENTIAL DEPENDENCIES

One of the most known expert finding methods based on the principle of co-occurrence is proposed by Balog et al. [1]. The method basically represents the propagation of relevance probability from documents to related candidates. It calculates the probability of expertness for a candidate as:

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

$$P(Q|D) = \prod_{q \in Q} P(q|D), \quad (2)$$

where $P(D)$ is distributed uniformly, $P(Q|D)$ is the probability of the document D to generate the query Q, measuring the document relevance according to the probabilistic language modeling principle of IR, $P(e|D)$ is the probability of association between the candidate and the document. The described method in fact models the assumption that the more often the candidate appears in documents containing many query terms the more likely the candidate is an expert. For that purpose, the method calculates the probability $P(e, q_1, ..., q_k)$ of observing the candidate expert e together with query terms $q_1, ..., q_k$ in some sample generated by language models of top documents, considering that the candidate and the query terms are generated independently given a ranked document.

Despite that the assumption of independence is very popular in various IR tasks, it does not always lead to the best performance. Including other features that characterize co-occurrence of terms often helps to improve. Two kinds of approaches are especially popular in document retrieval and, namely, in query expansion. While one approach measures the degree of proximity of a term to the query terms in the scope of a document [3], another one also takes the sequential order of terms into account [6]. While the assumption of independence between terms in a document is sometimes a too rough approximation, the assumption of independence between the document's content and its related persons seems even more doubtful.

Recently, it was shown for expert finding that the overall pairwise distance between a candidate's mention and the query terms in a document expresses the degree of association between the document and the candidate [7]. At the same time, the importance of an order in which personal identifiers and query terms occur in documents was never studied to the best of our knowledge.

We propose to use the sequential order of query terms and the candidate expert for estimating the amount of document relevance probability that should be propagated to the candidate. Since our approach is based on the assumption of dependency between query terms and the candidate, we rewrite Formula 1 in the following way:

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

$$P(Expert|e) = \sum_{D \in Top} P(Q|D)P(e|Q,D)P(D) \qquad (3)$$

$$P(e|Q,D) = \frac{a(e,Q,D)}{\sum_{e'} a(e',Q,D)}, \qquad (4)$$

where $P(e|Q,D)$ is the probability of association between the candidate and the document and $a(e,Q,D)$ is the non-normalized association score between the candidate and the document proportional to their strength of relation. Both *the probability and the association score are query-dependent.* They depend on where the candidate's personal identifier is mentioned in the text with respect to the positions of the query terms. We differentiate the following types of sequences to weight them differently:

- $a(e,Q,D) = w^{before}$: The candidate e is mentioned before any query term is mentioned $(e, q_1, ..., q_k)$,
- $a(e,Q,D) = w^{after}$: The candidate e is mentioned after all query terms are mentioned $(q_1, ..., q_k, e)$,
- $a(e,Q,D) = w^{between}$ The candidate e is mentioned in between of the query terms $(q_1, ..., e, ..., q_k)$.

Such orders may have various meanings, depending on specifics of a collection. For instance, authors of documents are usually mentioned before any topical words, people which are used to describe the topic probably occur somewhere in between of query terms and those who made lesser contributions are mentioned in the acknowledgments after all topical words. The proposed approach allows to distinguish these roles even when documents are not structured and persons mentioned in them are not semantically annotated. The experiments described in the next section simulate this widespread case.

3. EXPERIMENTS

We experiment with the **CSIRO** collection used by the Enterprise TREC community in 2007. It represents 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 extracted associations between candidate experts and documents by searching for the candidates email addresses and full names in the text of documents. It was enough to retrieve 50 documents containing at least one candidate's mention for the best performance of the baseline method.

Two expert finding methods are evaluated: the baseline method based on the assumption of independence between query terms and the candidate's mention (see Formula 1) and our method using the assumption of their sequential dependence (see Formula 3). For the baseline method the association score between the document and any candidate mentioned is always equal to 1.0. Since our method has only 3 parameters, we calculated their optimal setting with a simple coordinate-level hill climbing search method. The best performance of our method was reached with the following values for association scores: $w^{before}(e,Q,D) = 10.0$, $w^{after}(e,Q,D) = 0.1$, $w^{between}(e,Q,D) = 1.0$. In the case of several mentions of the same candidate in a document (what rarely happened), the maximum weight was used.

	MAP	MRR	P@5
Independence	0.361	0.508	0.220
Seq. Dependence	0.384	0.543	0.232

Table 1: The performance of expert finding methods using two assumptions: independence and sequential dependence of a candidate and query terms

This result suggests that the most important persons in the most relevant documents are always mentioned before topical words (for instance, it often happens in resumes that are very important evidence of expertness in CSIRO). Of course, this may vary between organizations and should actually depend on what kinds of documents prevail in the organization.

Both methods are compared in terms of common Information Retrieval performance measures officially used in TREC evaluations: Mean Average Precision (MAP), precision at top 5 ranked candidate experts (P@5) and Mean Reciprocal Rank (MRR). Table 1 demonstrates the advantage of our method based on the assumption of sequential dependence over the baseline method assuming sequential independence.

4. CONCLUSIONS AND FUTURE WORK

We presented an expert finding method taking the sequential dependence of a person and query terms occurring in a document into account. We claimed that it is useful to differentiate the orders of their occurrence in a document to estimate the strength of the relation of a candidate expert to the document's content. Experiments on the dataset from the Enterprise TREC 2007 justified our assumptions.

In the future we plan to develop a unified model combining all features describing co-occurrence into account: the proximity of a candidate to query terms, the order of their occurrence, the frequency of a candidate's occurrence etc. It is also promising to make the document-candidate weights dependent not only on the positions of candidate experts with respect to query terms positions in a document, but also on their absolute positions in the document or on the type of the document (at least, on its easily determined properties like extension or length).

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*, 2007.

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

[3] J. Gao, M. Zhou, J.-Y. Nie, H. He, and W. Chen. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *SIGIR'02*, 2002.

[4] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR'01*, 2001.

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

[6] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *SIGIR'07*, 2007.

[7] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *CIKM '07*, 2007.