Generative Modeling of Persons and Documents for Expert Search

Pavel Serdyukov, Djoerd Hiemstra, Maarten Fokkinga and Peter M.G. Apers
Database Group, University of Twente
PO Box 217, 7500 AE
Enschede, The Netherlands
{serdyukovpv, hiemstra, fokkinga, apers}@cs.utwente.nl

ABSTRACT

In this paper we address the task of automatically finding an expert within the organization, known as the expert search problem. We present the theoretically-based probabilistic algorithm which models retrieved documents as mixtures of expert candidate language models. Experiments show that our approach outperforms existing theoretically sound solutions.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: ;; H.3.3 [**Information Search and Retrieval**]: ;

General Terms

Algorithms, Theory, Performance, Experimentation

Keywords

expert finding, expertise, enterprise search, e-mail

1. INTRODUCTION

Expert finding is a new rapidly evolving direction of Information Retrieval research [4]. An expert search system finds persons with certain expertise within an organization. It uses a short user query and the information stored on personal desktops or within centralized databases as an input. There are mainly two approaches to do expert candidates modeling and ranking. The first approach is *profile-centric*. All documents related to a candidate expert are merged into a single personal profile. The personal profiles are ranked as in standard document retrieval and corresponding best candidates are returned to the user. The second approach is *document-centric*. It runs a query against all documents and ranks candidates by summarized scores of associated documents. Our generative modeling method combines the features of the both approaches: it ranks candidates using their language models built from retrieved documents.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

2. EXPERT MODELING

The most popular and effective assumption in expert finding research states that the level of personal expertise can be determined by the analysis of co-occurrence of the query terms and personal id within the context of a document or a passage [1]. We similarly suppose that our task comes to the estimation of the joint probability $P(e, q_1, ..., q_k)$ of observing the candidate expert e together with the query terms $q_1...q_k$ in the documents ranked by the query. In this paper we examine two estimation methods.

Method 1 considers a candidate and the query terms to be conditionally independent given a ranked document (see Figure 1a). Thus, the total joint probability is,

$$P(e, q_1, ..., q_k) = \sum_D P(D)P(e, q_1, ..., q_k|D) \quad (1)$$

$$P(e, q_1, ..., q_k|D) = P(e|D)\prod_{i=1}^{k} P(q_i|D)$$

This method is analogous to the most successful and theoretically sound approach proposed so far [1]. Thus, it serves as a baseline in our experiments. Method 2, which is the contribution of our paper, is based on the assumption of dependency between the query terms and a candidate. We suppose that candidates actually generate the query terms within retrieved documents (see Figure 1b). We calculate the required joint probability as follows considering the query terms to be sampled independently given an expert candidate:

$$P(e, q_1, ..., q_k) = P(q_1, ..., q_k|e)P(e) \quad (2)$$

$$P(q_1, ..., q_k|e) = \prod_{i=1}^{k} P(q_i|e)$$

We set the candidate prior $P(e)$ to be:

$$P(e) = \sum_D P(e|D)P(D) \quad (3)$$

So, now we need to estimate the probabilities $P(q_i|e)$. Since, we have already postulated that candidates are responsible for generating query terms in the documents they are mentioned in, we represent the language model of a ranked document as a mixture of expert candidate language

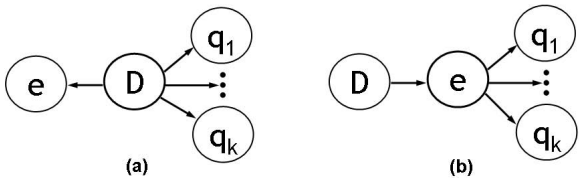

Figure 1: Dependence networks for two methods of estimating the joint probability $P(e, q_1, ..., q_k)$

models and the global language model. We define the likelihood of the top retrieved documents, set R, as:

$$P(R) = \prod_{D \in R} \prod_{w \in D} ((1-\lambda_G)(\sum_{i=1}^{m} P(e_i|D)P(w|e_i)) + \lambda_G P(w|G))^{c(w,D)}$$

Here, $e_1, ..., e_m$ are the experts, $c(w, D)$ is the count of term w in document D, λ_G is the probability that a term will be generated from the global language model and not from any of candidate language models. $P(w|G)$ is the global language model estimated over the whole document collection. Further, we apply the EM algorithm [3], traditionally used to estimate unknown parameters, to calculate $P(w|e_i)$. We propose the following updating formulas to be used recursively to maximize the likelihood of set R:

E-step:

$$P(e|w, D) = \frac{p(e|D)P(w|e)}{(1-\lambda_G)(\sum_{i=1}^{m} P(e_i|D)P(w|e_i)) + \lambda_G P(w|G)} \quad (4)$$

M-step:

$$P(w|e) = \frac{\sum_{D \in R} c(w, D)P(e|w, D)}{\sum_w \sum_{D \in R} c(w, D)P(e|w, D)} \quad (5)$$

The probabilities $P(e|D)$ are calculated using association scores $a(e, D)$ between the document and expert candidates: $P(e|D) = a(e, D)/\sum_{i=1}^{m} a(e_i, D)$. The probability distribution $P(D)$ is considered to be uniform in the both methods.

Our approach to expert candidates modeling is based on the similar hypothesis with one used in model-based pseudo-relevance feedback methods for document retrieval [5]. It considers that the relevance model of a user can be mined from the top of retrieved documents. The significant difference is that we represent the relevance model, which is in fact the model of the query topic, as a mixture of models of expert candidates who actually hold and share the desired knowledge.

3. EXPERIMENTS

For the evaluation of our approach we utilize data from the expert search task in the Enterprise track, TREC 2006. This track contains 1092 expert candidates and 50 queries with respective lists of experts. We use only the *email* part of the collection since it allows us to extract candidate-document associations easily, using *from*, *to* and *cc* email fields and given the candidate's email addresses. Association scores are taken to be 1.5, 1.0 and 2.5 respectively what is the best combination according to recent studies [2]. The number of retrieved documents for modeling is restricted to 1000 what is the standard for document retrieval tasks in TREC. The standard language model based IR approach is used for the retrieval of documents. Table 1 contains the results of both expert candidates ranking methods.

	MAP	MRR	R-pr	P5	P10	P20
M 1	0.1587	0.6550	0.2598	0.4285	0.4122	0.3341
M 2	0.1712	0.6712	0.2755	0.4306	0.4304	0.3653

Table 1: Performance of expert ranking methods

We see that our Method 2 improves the baseline Method 1 over all standard IR measures including mean average precision, mean reciprocal rank, R-Precision and precisions for 5, 10 and 20 top expert candidates. It shows that the assumption of independence of terms and candidates (see Figure 1a) associated with a document is less realistic. It seems important for expert ranking methods to model candidates, queries and documents considering that occurrence of the specific candidate in a document determines which terms the document consists of (see Figure 1b).

4. CONCLUSIONS

We presented a new generative model-based method for expert finding and evaluated it using the TREC Enterprise collection. The result suggests that it is more effective to drop the assumption of independence between candidates and document terms, while this claim should be carefully studied with additional experiments. We see the potential of the presented method for the understanding of expertise distribution in enterprise. In the future, we plan to take a closer look at pseudo-relevance techniques, namely query expansion, since our approach inherits much of this family of approaches.

5. ACKNOWLEDGMENTS

We thank Sergey Chernov, Gianluca Demartini and Julien Gaugaz from L3S Lab Hannover for the help with data preprocessing and series of fruitful discussions.

6. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, 2006.

[2] K. Balog and M. de Rijke. Finding experts and their details in e-mail corpora. In *15th International World Wide Web Conference (WWW2006)*, 2006.

[3] A. Dempster, N.M.Laird, and D.B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] D. Hawking. Challenges in enterprise search. In *ADC '04: Proceedings of the 15th Australasian database conference*, pages 15–24, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.

[5] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, 2001.