The Search for Expertise: to the Documents and Beyond

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

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, expertise search.

Expert finding is a rapidly developing Information Retrieval task and a popular research domain. The opportunity of search for knowledgeable people in the scope of an organization or world-wide is a feature which makes modern Enterprise search systems commercially successful and socially demanded. A number of efficient expert finding approaches was proposed recently. Despite that most of them are based on theoretically sound measures of expertness, they still use rather unrealistic and oversimplified principles. In our research we try to avoid these limitations and come up with models that go beyond the assumptions used in state-of-the-art expert finding methods.

The fundamental principle of existing approaches to expert finding is to infer expertise by analyzing the co-occurrence of personal identifiers and query terms in the scope of top ranked documents. While, the degree of co-occurrence of a person with topical terms is a reasonable evidence of personal expertness, the assumption about their independent occurrence seems not so adequate. In our methods, we consider that the occurrence of terms in the document is not independent from the presence of a candidate expert and vice versa. In one model, we regard people as generators of the expertise accumulated in the top retrieved documents. We extract their topic-specific personal language models that are further matched to a query [1]. In another model we simply assume that the responsibility of a person for the content of a document depends on its position in a document with respect to positions of the query terms and then just aggregate scores of documents related to a person for measuring personal expertness [3].

Suppose we still assume the independence of persons and terms in a document when measuring their co-occurrence. In this case we would in fact model the manual search for expertise by representing it as the following probabilistic process. 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 is 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. The described process can be interpreted as a *one-step relevance probability propagation* from documents to related candidate experts. However, the one-step probabilistic process is not quite a realistic model of a real-world user behavior. It is not likely that reading only one document and consulting only one person is enough to completely satisfy a personal information need in the enterprise. In our method, we also rely on the model of manual search for an expert, but in contrast to existing approaches, we do not assume that the user stops after the first step of moving from a selected document to the found candidate expert. We model *multi-step relevance probability propagation* from documents to candidate experts by means of K-step, infinite [2] or absorbing random walks [4]. We also show how we may benefit from adding direct organizational links among candidate experts.

Both approaches and the current state of expert finding research raise questions which we are going to address in our future work. Particularly, several promising directions are to be followed to extend the proposed solutions. While the assumption of dependence among persons and terms supposedly approximates the reality better, there are more ways to make it less straightforward. For instance, we may also assume the dependence between persons in a document. When we represent the expert finding task as a problem of ranking on graphs, some improvements specific to graph-based models seem rather appropriate. For example, new entities (e.g. dates, images, events) can be introduced into graphs in order to find new indirect associations among candidates and documents and hence better model the relevance flow.

REFERENCES

[1] P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *ECIR'08*, 2008.
[2] P. Serdyukov, H. Rode, and D. Hiemstra. University of Twente at the TREC 2007 Enterprise Track: Modeling relevance propagation for the expert search task. In *TREC'07*, 2007.
[3] P. Serdyukov, H. Rode, and D. Hiemstra. Exploiting sequential dependencies for expert finding. In *SIGIR'08*, 2008.
[4] P. Serdyukov, H. Rode, and D. Hiemstra. Modeling expert finding as an absorbing random walk. In *SIGIR'08*, 2008.

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