Lightweight Semantic Search Based on Heterogeneous Sources of Information

Marián ŠIMKO*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
simko@fiit.stuba.sk

To satisfy user's information needs, the most accurate results for entered search query need to be returned. Traditional approaches based on query and resource Bag-Of-Words model comparison are overcome. In order to yield better search results, the role of semantic search is increasing. However, the presence of semantic data is not common as much as it is needed for search improvement [1]. Although there are initiatives to make resources on the Web semantically richer, it is demanding to appropriately describe (annotate) each single piece of resource manually. Furthermore, it is almost impossible to make it coherently. The current major problem of the semantic search is the lack of available semantics for the resources, especially when considering the search on the Web [2].

To overcome this drawback, we propose an approach leveraging lightweight semantics of resources. It relies on resource metadata model representing resource content. It consists of interlinked concepts and relationships connecting concepts to resources (subjects of the search) or concepts themselves. Concepts feature domain knowledge elements (e.g. keywords or tags) related to the resource content (e.g. web pages or documents). Both resource-to-concept and concept-to-concept relationship types are weighted. Weights determine the degree of concept relatedness to resource or other concept, respectively. The domain representation we obtain is straightforward and adopted to its goal – it is designed to addresses the specifics of the Web environment and enables to improve personalized search. Furthermore, it resembles lightweight ontology thus allowing automated generation.

When acquiring metadata, we process heterogeneous sources of information: the content and social data. We consider:

- keywords supplied by the author himself,
- keywords (concepts) generated automatically by content processing,

* Supervisor: Mária Bieliková, Institute of Informatics and Software Engineering

Spring 2010 PeWe Workshop, April 18, 2010, pp. 19–20.

- tags supplied by both web-application users or users of some social tagging service.

We assume that by using heterogeneous sources of information the acquired domain model will be more accurate and thus feasible to enable advanced behavior such as recommendation or personalized search in web-based applications.

Having domain model as described above, we examine the possibilities of search improvement. We propose two variants of so called *concept scoring computation* taking place online during searching (see Figure 1). With concept scoring we extend the baseline state-of-the-art approaches to query scoring computation expecting an improvement of the search. For the computation we consider two approaches: statistical and topological. First approach takes into account statistical aspect of available metadata, while second one analyses a subset of metadata topology. Utilizing metadata we are able to assign the query to particular topic (set of concepts) and yield more accurate search results with respect to related resources.

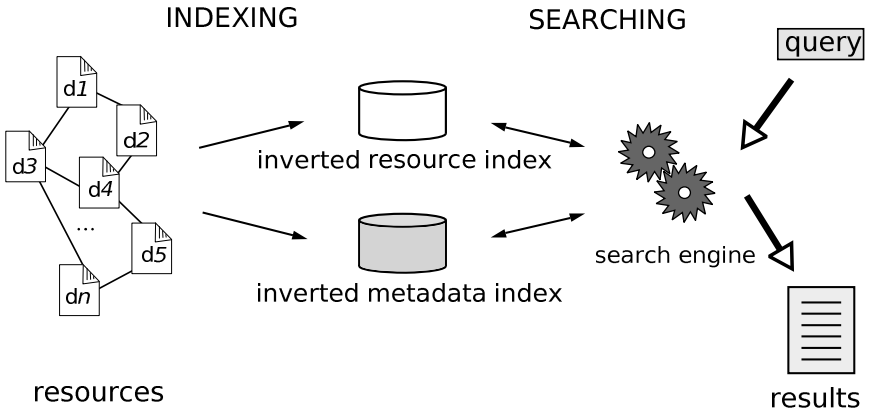

Figure 1. Combined scoring computation overview.

In the current stage of the research we are working on the evaluation of the proposed approach by building on the Lucene information retrieval library.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0508/09.

References

[1] Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E., Castells, P. Semantic Search Meets the Web. In Proc. of the 2008 IEEE International Conference on Semantic Computing, ICSC 2008, pp. 253–260 (2008)

[2] Sabou, M., Gracia, J, Angeletou, S., d'Aquin, M., Motta, E. Evaluating the Semantic Web: A Task-based Approach. In LNCS 4825: The Semantic Web. Proc. of the 6th International Semantic Web Conference, ISWC 2007, Busan, Korea, pp. 423–437 (2007)