Improving Search Results with Lightweight Semantic Search

Marián Šimko
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technology, Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
simko@fiit.stuba.sk

Mária Bieliková
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technology, Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
bielik@fiit.stuba.sk

ABSTRACT

The goal of each search service is to yield the most relevant results on a given query. Traditional full-text search is not enough and many approaches to improve search rankings are adopted. In this paper we propose a method of combined search query scoring computation leveraging lightweight semantics represented by metadata related to searchable content. It extends state-of-the-art approaches at both indexing and searching stage. We discuss two approaches of so-called *concept scoring* computation in order to capture different properties of available metadata.

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

General Terms: Algorithms, Design, Experimentation

Keywords: semantic search, folksonomy, concept scoring

1. INTRODUCTION

To satisfy user's information needs, the most accurate results for entered search query need to be returned. Traditional approaches based on the query and resource bag-of-words model comparison are obsolete [2]. In order to yield better search results, the importance of semantic search is increasing. However, the presence of semantics is not common as much as it is needed for search improvement. Although several initiatives to make resources on the Web semantically richer exist (e.g. [5]), it is demanding to describe (annotate) appropriately 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.

In this paper we introduce an approach relying on *lightweight* semantics referred to as the resource *metadata*. Resource metadata represent a simplified semantic model of the resources. 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* relationships are weighted. Weights determine the degree of concept relatedness to the resource or to other concept, respectively. Interlinked concepts result in a structure resembling lightweight ontology and form a layer above the resources allowing an improvement of the search.

The advantage of modeling domain knowledge as described above lies in its simplicity. Hence, it is possible to generate metadata automatically enabling lightweight semantic search for a vast majority of resources. We have already performed several experiments of automatic metadata extraction with promising results in e-learning domain [4]. In addition, existing and evolving folksonomies can supplement extracted metadata.

We propose the method of concept scoring computation which utilizes concept relations associated to resources matching the query. Having this information we are able to assign the query to the particular topic (set of concepts) and restrict search results only to related resources. The whole method is set into the traditional search context and is divided into the following steps (see Figure 1):

1. Indexing (offline):
   (a) inverted resources index composition,
   (b) inverted metadata index composition;
2. Searching (online for each query):
   (a) basic scoring computation,
   (b) concept scoring computation,
   (c) scoring combination.

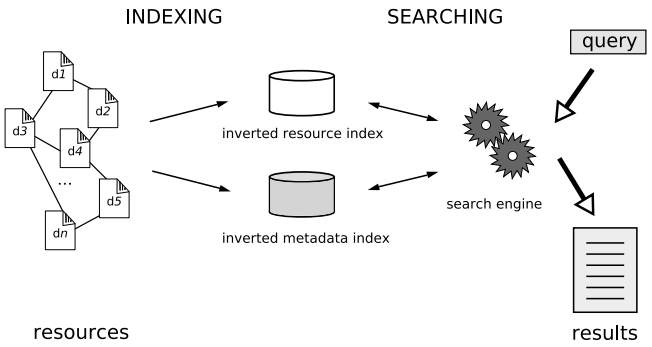

Figure 1: Combined scoring computation overview.

While following traditional search paradigm, we present a novel approaches to both metadata indexing and query-time scoring computation (steps 1b and 2b) and combine them with baseline state-of-the-art approaches (steps 1a and 2a).

WWW2009, April 20-24, 2009, Madrid, Spain.
.

2. INVERTED METADATA INDEX

When adding a resource to the collection, besides creating the representation of its content we additionally compose a metadata model [3]. First, we extract related concepts. Then we compute resource-to-concept relatedness weights. Finally, employing a link analysis, we discover and create relationships between concepts.

The methods for metadata generation (concepts extraction) can vary. In our experiments we extracted concepts as the most relevant terms from document representations including domain keywords available in the resource content. Relations between concepts were generated by spreading activation over actual metadata graph (consisting of interlinked documents and concepts). In addition to resource content processing, as concept candidates we can also consider tags provided collaboratively by users.

When storing metadata, we actually benefit from simplicity of domain knowledge modeling. Viewing metadata as a graph, it is easy to store its content into the inverted metadata index. The index maps both documents and concepts onto interconnected concepts in metadata, containing also a weight of the relations (Table 1).

Table 1: Inverted metadata index example.

Resource	Concepts
R1	(polymorphism,0.17), (inheritance,0.12)
R8	(index,0.59), (array,0.11), (for,0.02)
R9	(stream,0.10), (save,0.61), (output,0.2)
stream	(save,0.8), (read,0.6), (pipe,0.41)

3. CONCEPT SCORING

Having indices prepared, we are able to compute the query scoring. After entering the query, we first compute "basic" score as it is currently done in common full-text search engines. Basic score is derived using inverted text index and is typically based on well-known tf-idf weighting. We additionally compute the concept score, which utilizes inverted metadata index and boosts documents' scorings related to the set of concepts for which the query is typical. This way we identify the context of query by traversing the metadata graph. Scoring computation is extended from an occurrence analysis to lightweight classification task.

For the computation we consider two approaches: statistical and topological. *Statistical* concept scoring computation is based on well-known tf-idf analogy: the more concepts the document is related to exist, the less relevant is to the query. We first identify all concepts (referred to as *topic*) all documents matching query phrase are connected with. The statistical concept scoring will boost documents more related to the topic.

Topological approach to the concept scoring computation is in its core a link analysis approach. After the topic identification the prestige (implicit node ranking obtained using advanced PageRank [6]) of the concepts within is computed. The document's topological concept scoring is computed as weighted sum of resource-to-concept relation weights:

$$s_c^t(q,d) = \sum_{c \in T} r_{d,c} p'_c \qquad (1)$$

where s_c^t is topological concept scoring, topic T is set of all concepts related to query q, $r_{d,c}$ is weight between document d and concept c and p'_c is normalized prestige of concept c. Here, the biggest influence on results has the most "dominant" concept within the topic according to its prestige.

4. CONCLUSIONS

We have presented an approach to semantic search leveraging metadata available for the resources. Metadata have simple structure, thus can be acquired automatically by preprocessing the resources. Furthermore, metadata can be derived from tags for the resources (folksonomies), defined collaboratively using some of the Semantic Web services.

In presented method the query scoring is computed as a combination of traditional approaches based on the full-text search with a novel part – concept scoring computation – based on metadata available for the searched content. Automatic extraction of concepts is a crucial part especially in large and dynamic information spaces such as the Web. Our experiments in the e-learning domain showed that extracted concepts were capable of identifying resources and thus serve for recommendation.

We proposed two methods of the concept scoring: statistical and topological. The advantage of the first one is its computational complexity, which do not exceed state-of-the-art approaches. The computational complexity of the second one is higher due to the link analysis of metadata, which is performed online. Currently we experiment with this part of the method to justify our proposal.

Main advantage of proposed approach lies in actual availability of metadata. We believe that considering large metadata space interconnected with resources space can improve search in general more than small non-interlinked islands of semantics. The combination of the approaches can bring even better results.

A further advantage of the proposed method is the possibility of the combination with other search algorithms (for example those based on link analysis between documents [1]) and thus yielding even more accurate search results.

Acknowledgement

This work was partially supported by the Cultural and Educational Grant Agency SR, grant No. KEGA 3/5187/07 and by the Scientific Grant Agency SR, grant No. VG1/0508/09.

5. REFERENCES

[1] Brin, S., Page, L. The anatomy of a large-scale hypertextual web search engine. In Proc. of the 7th Int. World Wide Web Conf., 1998.

[2] Haveliwala, T. Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search. In IEEE Transactions on Knowledge and Data Engineering 15(4), 2003, pp. 784-796.

[3] Šimko, M., Bieliková, M. Domain concepts relationships discovery (In Slovak). In Proc. of 3rd Workshop on Intelligent and Knowledge oriented Technologies, WIKT 2008, Smolenice, Slovakia, pp. 17-20.

[4] Šimko, M., Bieliková, M. (Semi)Automatic e-course metadata generation (In Slovak). In Proc. of Znalosti 2009, Brno, Czech republic, 2009, pp. 246-257.

[5] Tran, D., et al. Expressive Resource Descriptions for Ontology-Based Information Retrieval. In Proc. of the 1st Int. Conf. on the Theory of Information Retrieval (ICTIR'07), 2007, Budapest, Hungary, pp. 55-68.

[6] White, S., Smith, P. Algorithms for estimating relative importance in networks. In Proc. of the ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data mining. ACM Press, 2003, pp. 266-275.