Towards Mapping-Based Document Retrieval in Heterogeneous Digital Libraries

Heiner Stuckenschmidt[1], Wolf Siberski[2], Erik van Mulligen[3,4]

[1] Vrije Universiteit Amsterdam
[2] L3S Research Center, Hannover
[3] Erasmus University Rotterdam
[4] Collexis BV, Geldermalsen

June 28, 2005

In many scientific domains, researchers depend on a timely and efficient access to available publications in their particular area. The increasing availability of publications in electronic form via digital libraries is a reaction to this need. A remaining problem is the fact that the pool of all available publications is distributed between different libraries. In order to increase the availability of information, these different libraries should be linked in such a way, that all the information is available via any one of them. Peer-to-peer technologies provide sophisticated solutions for this kind of loose integration of information sources. In our work, we consider digital libraries that organize documents according to a dedicated classification hierarchy or provide access to information on the basis of a thesaurus. These kinds of access mechanisms have proven to increase the retrieval result and are therefore widely used. On the other hand, this causes new problems as different sources will use different classifications and thesauri to organize information. This means, that we have to be able to mediate between these different structures. Integrating this mediation into the information retrieval process is a problem that to the best of our knowledge has not been addressed before.

Recently a number of approaches have been proposed for the purpose of matching heterogeneous classifications based on different criteria such as lexical matches between classes, structural similarity of the classification hierarchy or logical definitions of classes. A common feature of most of these approaches is their inability to determine exact matches. They rather determine the degree in which two concepts are similar. This measure of similarity provides us with a way of integrating mappings in to the information retrieval process. The idea is to determine the relevance of a document not only based on the documents score which is calculated using standard information retrieval measures such as

Dagstuhl Seminar Proceedings 05271
Semantic Grid: The Convergence of Technologies
http://drops.dagstuhl.de/opus/volltexte/2005/385

TFxIDF, but to combine this with a concept score that assesses the similarity of the concept a document is assigned to a query concept. The combined score can be determined by the weighted sum of the two measures. Given a search term t_1, the combined relevance of a document d can be computed according to the following formula.

$$\alpha \cdot CS(t_1, t_2) + \beta \cdot DS(d, t_2), \text{where } \alpha + \beta = 1$$

In most cases, users are only interested in documents that best match their query. This observation can be exploited for providing efficient retrieval strategies that return the k top ranked documents. These techniques optimize the number of documents to be checked using some heuristics (e.g. only the top k documents of each source have to be considered, at most k sources will provide documents). Similar heuristics can be applied in our setting to provide efficient methods for retrieving the k most relevant documents from distributed digital libraries. For instance, only documents assigned to concepts with a certain level of similarity to the query concept have to be considered. As in most cases, the number of concepts will be orders of magnitude smaller that the number of documents, this can lead to a significant reduction of retrieval time compared to a brute force approach where all documents are tested for their relevance.

We have set up an experiment to test the idea of mapping-based information retrieval. The experiments are built on top of the Collexis document indexing and retrieval software and uses a standard benchmark Dataset from the medical area. The dataset consists of 300.000 articles from medical journals. The articles are annotated with relevant terms from the MeSH thesaurus.

Figure 1 illustrates the general setting of the experiments which consist of three steps:

1. As a first step, we will distribute the documents in the OHSUMED dataset to simulate a collection of distributed digital libraries. Each library will correspond to one sub-vocabulary of UMLS and be represented by a peer in a P2P system.

2. In a second step, we will conduct retrieval experiments based on the combines measure introduced above to retrieve relevant documents from all peers using the different sub-vocabularies

3. The third step will be a comparison of the results of step two with an approach that only uses the document score of the search term in order to find out whether we gain or loose anything by taking mappings in to account.

For this experiment we use the Collexis engine to automatically assign documents to terms in the UMLS metathesaurus. This process results in a so-called

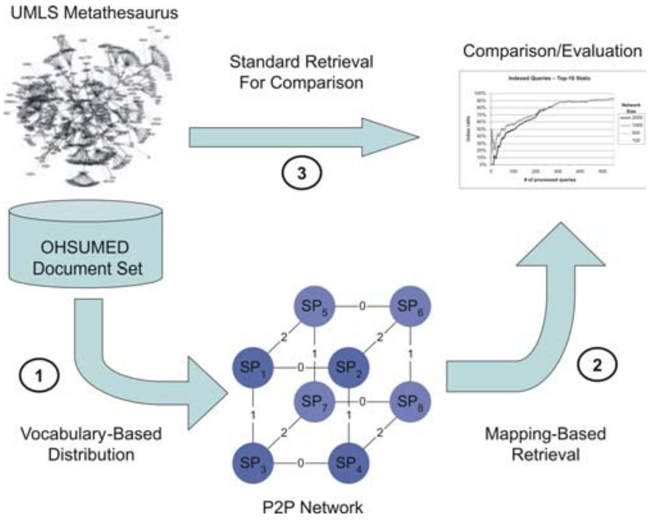

Figure 1: General Setting of the Retrieval Experiments

fingerprint for each document that contains a list of concepts and the degree to which the concept is relevant for the document. We use this degree of relevance as our document score $DS(d, t_2)$. The Collexis system uses the normalized term frequency for determining this score:

$$DS(d,t) = \frac{\mathbf{tf}_{t,d}}{max_i\{\mathbf{tf}_{i,d}\}} \cdot log\left(\frac{D}{\mathbf{df}_t}\right)$$

where $\mathbf{tf}_{t,d}$ is the number of occurrences of the term t in document d and $\mathbf{df}_t$ is the number of documents that contain the term t.

As UMLS has been created by merging more than 50 independent medical terminologies, it can be used to simulate the use of different classifications. In particular, the metathesaurus contains all the terms from the integrated terminologies and links them using standard thesaurus relations such as synonym-of, broader-term and narrower term. We can use this integrated structure to determine the similarity of concepts from different terminologies in terms of their semantic distance in UMLS. A standard measure for semantic similarity is the following:

$$CS(t_1,t_2) = \begin{cases} e^{-\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} & \text{if } t_1 \neq t_2, \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

where l is the length of the shortest path between topic t_1 and t_2 in the hierarchy, h is the level of the direct common subsumer of t_1 and t_2 and α and β are scaling factors that determine the contribution of the shortest path and the depth of the hierarchy, respectively.

In the experiments outlined above, we will make a first step towards mapping-based information retrieval across heterogeneous classifications. There are many ways in which this basic setting can be extended to better fit real-world requirements. Three extensions are of particular interest for our work. The first is an extension of the approach to multiple search terms. While for the document score, there are standard ways of extending the measure to multiple search terms that are already implemented in the Collexis System, we have to determine how the concept relevance measure is determine for multiple concepts and how it is combined with the document score. The second extension concerns the use of existing mapping algorithms for determining the similarity independent of the UMLS meta-thesaurus. This does not only make the approach applicable in other domains as well, it also provides a new approach for evaluating mapping algorithms based on the quality of the retrieval results. The third extension is to apply the approach in a distributed setting where peers provide information based on different terminologies hosted at super-peers. The challenge of such a setting is the need to provide query routing methods that minimize the communication costs between different peers in the system and still deliver complete query results.