Improving Biomedical Document Retrieval using Domain Knowledge

Shuguang Wang
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
swang@cs.pitt.edu

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
milos@cs.pitt.edu

ABSTRACT

Research articles typically introduce new results or findings and relate them to knowledge entities of immediate relevance. However, a large body of context knowledge related to the results is often not explicitly mentioned in the article. To overcome this limitation the state-of-the-art information retrieval approaches rely on the latent semantic analysis in which terms in articles are projected to a lower dimensional latent space and best possible matches in this space are identified. However, this approach may not perform well enough if the number of explicit knowledge entities in the articles is too small compared to the amount of knowledge in the domain. We address the problem by exploiting a domain knowledge layer, a rich network of relations among knowledge entities in the domain extracted from a large corpus of documents. The knowledge layer supplies the context knowledge that lets us relate different knowledge entities and hence improve the information retrieval performance. We develop and study a new framework for i) learning and aggregating the relations in the knowledge layer from the literature corpus; ii) and exploiting these relations to improve document retrieval.

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

General Terms: Algorithm, Performance

Keywords: Information Retrieval, Link Analysis, Biomedical Documents

1. INTRODUCTION

Due to the complexity of the scientific domains today, research documents may feasibly mention only a fraction of knowledge of the field. This is not a problem for humans who are armed with a general domain knowledge and hence are able to overcome the missing links and connect the information in the article to the overall body of domain knowledge. But many search and information-retrieval systems that work by analyzing and matching queries only to individual documents are very likely to miss these knowledge-based connections and thus fail to retrieve many relevant documents.

The objective of our work is to improve the retrieval of relevant documents using a knowledge layer, a rich network of connections relating knowledge entities in the domain. The relations in the knowledge layer are build automatically by mining associations in large corpora of documents. The intuition is that document level analysis used for example in PLSI[2] is not sufficient to extract models rich enough to overcome the complexity of the domains and relative sparseness of knowledge entities referenced in individual documents. Globally extracted and aggregated relations can be then used to infer hidden but closely related knowledge entities that are relevant for the query. The advantage of our approach and its inference abilities is that it can be combined easily with existing information retrieval techniques. We demonstrate this by combining the method with PLSI[2] and BM25[3] retrievals.

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

2. THE APPROACH

The framework proposed in this work (1) extracts a knowledge model from scientific documents in large domain corpora, (2) uses the model to support inferences on domain entities, hence improving the retrieval of relevant documents.

2.1 Extracting Domain Knowledge Model

The knowledge of any scientific filed can be seen as a rich network of relations among domain entities. Due to the complexity of scientific domains it is infeasible to mine all these relations from a single document, a more complete picture arises only if the information in a large corpora is aggregated. However, it is still unclear how one can use the network of relations extracted from many documents to support inferences for the information retrieval purposes. We propose to study link analysis methods to address the problem. Our intuition is that closely and tightly related knowledge entities are more relevant to each other.

To perform link analysis we rely on PHITS [1]. PHITS is a probabilistic model used to study document citation and web hyperlink structure. Mathematically it is similar to PLSI[2]. PHITS assumes a naive bayes decomposition of documents on a latent variable z representing different communities of documents. In our work, we use it to model relations among domain entities (terms), and not documents. We assume that each entity e_i is independent of others given a latent factor z, which represents a family of closely related domain entities. PHITS lets us represent the relations among entities indirectly using the relations between entities and latent variables. These relations are represented as conditional probabilities $P(e_i|z)$ and the full joint distribution

over all entities is defined as:

$$P(e_1, \ldots, e_i, \ldots, e_n) = \sum_z \prod_{i=1}^{n} P(e_i|z)p(z) \quad (1)$$

2.2 Inferences on Domain Knowledge

Given the PHITS-based model we can define probabilistic relations between entities and documents as well. Each document d_i is represented as a vector of domain entities that occur in it. Zero entries in the vector means some domain entities are not mentioned in the document. However, this does not mean they are irrelevant to the document. Instead we treat them as hidden variables and our ultimate goal is to figure out probable values of these hidden variables.

$$\begin{aligned} P(e_h|d_i) &= \sum_z P(e_h|z, M_{phits})P(z|e_{d_i,1}, e_{d_i,2}, \ldots, e_{d_i,k}, M_{phits}) \\ &\sim \sum_z P(e_h|z, M_{phits})p(z) \prod_{j=1}^{k} P(e_{d_i,j}|z, M_{phits}) \end{aligned} \quad (2)$$

Equation 2 explains how to compute the probability of of a hidden entity e_h in document d_i containing entities $e_{d_i,j}$ with the help of the PHITS model.

3. EVALUATION

The relevance of a scientific document to the query is best assessed by a human expert. Unfortunately this can be a very tedious and costly process. To alleviate this problem and still demonstrate the benefit of our approach we use the following setup: we perform all knowledge-model learning and retrieval analysis on documents' abstracts only; full texts and exact matches of queries on full texts serve as surrogate measures of relevance. Briefly, if we retrieve a document based on its abstract, the relevance of the abstract is judged based on the exact match of the query to the full document.

We evaluate our approach on a corpus of 6000 cancer study articles from PubMed. Domain-specific entities considered in our analysis are names of genes and proteins and we used BIC[4] to choose the best number of latent factors. To demonstrate the benefit of our method and the knowledge layer we combine our method with PLSI[2] and BM25[3]. In addition, we compare the results to Lucene and run it once on only abstracts, the other time on full text.

We run two sets of experiments and use the results of Lucene indexed with abstracts as the baseline. First, we test the methods on 500 queries that involve random pairs of genes and proteins. Figure 1 shows the 11-point average precision of different methods. It is not surprising that Lucene indexed with full text performs the best. Original PLSI and BM25 do not outperform the baseline. However, they both do better when combined with our methods and domain knowledge. The results clearly show that domain knowledge helps to improve the retrieval.

In second experiment, we compare different approaches with another type of queries. Besides random queries, we construct some of the queries from the species names occurs only in main body of the testing documents. We call them content queries. Thus it is unlikely to match them in abstracts and they can demonstrate if the mined domain knowledge helps to provide the absent context knowledge of domain entities. Table 1 shows the interpolated average precision of different systems with various portion of content queries.

Overall, we improve BM25 and PLSI with domain knowledge and they both perform better than the baseline. In most cases, the relative improvement of our approaches increases as we have more content queries. This indicates that mined domain knowledge provides the missing context information in the abstracts.

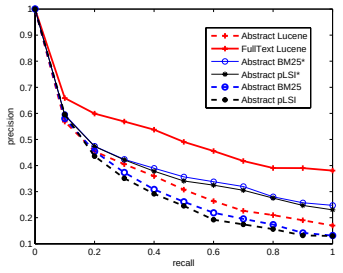

Figure 1: 11-Point Interpolated Average Precision

Table 1: Interpolated average precision for different query combinations ('Abs' - Abstract, 'Full' - Full-Text, '*' - Combined Approaches)

	10%	30%	50%	80%	100%
Abs *Lucene*	0.37	0.28	0.21	0.16	0.11
Full *Lucene*	+45%	+75%	+104%	+137%	+200%
Abs *BM25**	+10%	+11%	+14%	+12%	+17%
Abs *PLSI**	+2%	+7%	+4%	+6%	+9%
Abs *BM25*	−5%	−3%	−0%	−0%	−0%
Abs *PLSI*	−10%	−10%	−14%	−12%	−9%

4. CONCLUSION AND FUTURE WORK

We present a framework that extracts the domain knowledge from multiple documents and uses it to support document retrieval inferences. We have shown that our method can improve the retrieval performance on biomedical literature. Recent research projects like [5] also use domain knowledge and demonstrate its benefit. However, to our knowledge this is the first work that attempts to learn the probabilistic relations among domain entities and use them in document retrieval. In future, we would like to verify our framework on retrieval of full text documents and build a more comprehensive domain knowledge that explicitly captures a variety of domain relations.

5. REFERENCES

[1] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *ICML 2000*.

[2] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR 1999*.

[3] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *TREC 1994*.

[4] G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[5] W. Zhou, C. Yu, N. Smalheiser, V. Torvik, and J. Hong. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR 2007*.