Improving Biomedical Document Retrieval by Mining Domain Knowledge

Shuguang Wang
Intelligent Systems Program
University of Pittsburgh

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh

Abstract

When research articles introduce new findings or concepts they typically relate them only to knowledge and domain concepts of immediate relevance. However, many domain concepts relevant for the article and its findings are omitted in the text. This may prevent us from retrieving articles of interest when executing a search query. Approaches such as probabilistic latent semantic indexing (PLSI) overcome this limitation by projecting terms in articles to a lower dimensional latent space and best possible matches in this space are identified. Nevertheless, this approach may not perform well enough if the number of explicit knowledge concepts in the articles is too small compared to the amount of knowledge in the domain. The objective of this paper is to address the problem by exploiting a domain knowledge layer: a rich network of associations among knowledge concepts in the domain of interest. We present a new document retrieval framework that i) extracts associations among knowledge concepts from many documents in the literature corpus; ii) and exploits them to improve the retrieval of relevant documents. We test our approach on the problem of retrieval of biomedical documents and show that it outperforms standard Lucene and BM25 information-retrieval methods.

Introduction

A huge number of research articles have been published in different areas of science and thousands of new papers are added every year. With the growth of the scientific knowledge the finding of the information the researchers are interested in is becoming an increasingly hard task. In order to explore some topic or a hypothesis in the domain, the researchers may rely on general web search engines such as Google, Yahoo!, or various journal reference databases like PubMed (www.ncbi.nlm.nih.gov/pubmed) that return the documents that match the query keywords issued by the researchers. However, because of 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 knowledge of the field and hence are able to overcome the missing link and connect the information in the article to the overall body of domain knowledge. Nevertheless many existing 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. Hence many documents that are extremely relevant for the query may not be returned by the existing search systems.

The goal of our research is to study the influence of knowledge on information retrieval and its ability to find better, more relevant documents when they do not exactly match the search query. We present a new document retrieval framework that, similarly to humans, attempts to exploit domain knowledge to strengthen the retrieval process. The knowledge model proposed in this work has a fairly simple form and consists of a collection of pair-wise associations among domain concepts. However, we note that despite its representational simplicity, the association networks for complex domains with thousands of domain concepts can become fairly complex.

In general, associations may stand for and abstract a variety of relations among domain concepts. One reason for using the associations instead of complex relational models is that these are relatively easy to mine from the text, hence the model can be build automatically from a large corpora of domain documents. The second reason is that association networks and patterns therein give clues about mutual relevancy of knowledge concepts. Our hypothesis and that highly interconnected knowledge concepts define semantically relevant groups, and that these patterns can be used to perform useful information-retrieval inferences, such as those connecting hidden and explicitly mentioned knowledge concepts in the document.

The analysis of network structures and patterns therein is typically conducted using link analysis methods. We propose to use PHITS framework (Cohn & Chang 2000) to analyze the mutual connectivity of knowledge concepts in association networks and derive probabilistic relations we expect, if our hypothesis is correct, to reflect the mutual relevance of the domain concepts. We note that this is very different from typical PHITS applications that analyze links in the co-citation networks or a web hyperlink structure. In particular, these applications attempt to understand the structures connecting documents, while in our case we want to analyze the relation among domain concepts in the knowledge model.

We experiment with and demonstrate the potential of our

framework on documents in the biomedical literature using search queries on protein and gene species referenced in these articles. The knowledge model used for information retrieval is composed of relations among these domain concepts. Our results show that the addition of knowledge layer inferences improves the retrieval of relevant documents and outperforms the state-of-the-art retrieval systems based on Lucene (lucene.apache.org), PLSI(Hofmann 1999), and BM25(Robertson *et al.* 1994).

Methods

Our objective is to improve the information retrieval performance of scientific documents with the help of a knowledge model that relates domain-specific concepts. Examples of domain-specific concepts are names of genes, proteins or diseases in the biomedical literature.

The framework proposed in this work (1) extracts the domain knowledge from the documents in large domain corpora; (2) uses it to support inferences on related domain concepts; and (3) feeds the inference results into other information retrieval procedures.

Domain Knowledge Model

The knowledge of any scientific field can be seen as a rich network of relations among domain concepts. If the knowledge is represented in a computer, it can be used to support inferences on related knowledge concepts and subsequently improve the information retrieval performance.

The domain knowledge can be supplied into the information retrieval system directly by a human expert, or indirectly through publicly available resources, for example, published domain ontologies such as those defined in OpenCyc (www.opencyc.org) or GO(Consortium 2000). In this work we adopt a different strategy. We assume that no prior knowledge source is available and the knowledge model useful for retrieval is built solely from existing document collections.

Domain concepts extraction Domain-specific concepts considered in our analysis are names of genes and proteins. The species names in the articles were extracted with the help of the naming program we built. Briefly, the program performs the following steps:

1. Abstracts are segmented into sentences.
2. Each sentence is tagged using a POS tagger (MedPost) developed by NCBI for biomedical literature.
3. Tagged sentences are parsed using Collin's full parser(Collins 1999).
4. The phrases are matched with the species names based on the GPSDB(Pillet *et al.* 2005) vocabulary database.
5. The synonyms of species are matched and each distinct species is assigned a unique identification.

This naming program is able to achieve over 90% precision at about 65% recall when extracting gene and protein species name on a 100 document testing set. In our study, we are concerned more about the precision of the program as we want to introduce least false information possible. Although we do not extract all species in the documents, we are still able to learn interesting relations and demonstrate the benefit in document retrieval in the evaluation.

Association network The knowledge we rely on in our work has a relatively simple form and consists of a network of associations in between domain concepts. The associations correspond to pairs of concepts referenced in the same sentence, the same sentence group, or in the same paragraph, depending on how accurate we would like to be when extracting these patterns. In other words, if the two concepts appear in the same sentence (or a paragraph) the association in between the two concepts is established and included in the network. The association network is built by aggregating and merging associations relations from a large corpus of document. If a pair of species co-occur in multiple documents, we count the pair only once. The motivation for building the model this way is that pieces of domain knowledge are scattered among many documents, and more complete knowledge arises only if the pieces are aggregated. The advantage of the association-based domain knowledge model is that due to its simplicity it is easy to automatically mine from the text.

We are interested in turning the above knowledge model into a method that can support inferences on terms (concepts) that are relevant to the terms in the query. However, it is still unclear how the association network may define the relevancy among terms. We propose to model mutual relevancy indirectly using their interconnectedness in the association network. More specifically, our hypothesis is that domain concepts are more likely to be relevant to each other if they belong to the same, well defined, and highly interconnected group of concepts. The intuition for our approach is that concepts semantically interconnected in terms of their roles or functions should be more relevant to each other. At the same time we expect this semantically distinct roles and functions are reflected in the documents and hence picked up and reflected in our association network.

To explore and understand the interconnectedness of domain concepts in the association network we resort to *link analysis methods*. Intuitively, the concepts are related if they belong to a well defined group of highly interconnected concepts. We rely on PHITS model (Cohn & Chang 2000) to perform the link analysis.

Probabilistic HITS PHITS (Cohn & Chang 2000) is a probabilistic link analysis model that was used to study graphs of co-citation networks or web hyperlink structures. However, in our work, we use it to study the association network among domain concepts (terms) and not documents. This is very important distinction to stress, since the PHITS model on the document level has been used to improve the search and information retrieval performance as well (Cohn & Chang 2000). The novelty of our work is to use PHITS to assess the mutual relevancy of domain concepts.

The PHITS model is learned from the link structure data using the maximum likelihood criterion. To do this PHITS relies on a tempered Expectation-Maximization (EM) approach (Hofmann 1999). In the expectation step, it computes the expectation that a pair of concepts is "related"

by a latent factor z, i.e., $P(z|e_i, e_j)$. In the maximization step, it re-estimates parameters defining distributions $P(z)$, $P(e_j|z)$, and $P(e_i|z)$. More details about the specifics of the tempered EM can be found in PHITS paper (Cohn & Chang 2000).

PHITS models with different number of latent latent factors are possible. To decide on what model to use various model selection methods can be applied. However, instead of using just one model we resort to model averaging where the results of inference from different models are averaged. In context of information retrieval, the model averaging approach has been attempted and shown helpful in (Wei & Croft 2006).

Inferences on Domain Knowledge

Given the PHITS-based model we can define probabilistic relations between concepts and documents. Each document d_i is represented by domain concepts occurring in the document. However, this does not mean unobserved (absent) concepts are irrelevant to the document. A subset of them, especially those that are closely and tightly related to the explicitly mentioned concepts, may be very relevant to the document. The strength of this relation is captured probabilistically by the PHITS model.

To assess the relevance of a concept e_h to document d_i we use PHITS to approximate the probability of the concept being 'generated' by the document in the next step as:

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

where $e_{d_i,1}, \ldots e_{d_i,k}$ are concepts explicitly observed in the document d_i. M_{phits} is the PHITS model we learned. The equation (when applied to all possible concepts) defines a distribution over concepts that reflects how likely different concepts are to be generated next by the given document.

Using Domain Knowledge in Information Retrieval

Our objective now is to use PHITS and Equation 1 to transform the term (concept) vector representing a document into a new term vector in which all unknown (unobserved) concepts are represented by their inferred values. Figure 1 illustrates the idea and contrasts it to typical information retrieval process pipeline in which terms not observed in the document are left intact.

The caveat of using the PHITS model to infer probabilities of all unknown concepts is that PHITS treats all concepts as alternatives. That is, the probability calculated in Equation 1 does not represent the probability $P(e_h = T|d_i)$ with which a concept e_h is expected to occur in the document. To resolve the problem we approximate this probability as:

$$P(e_h = T|d_i) = \min[\alpha * P(e_h|d_i), 1] \quad (2)$$

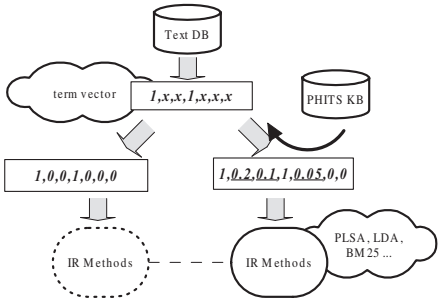


Figure 1: Exploitation of the domain knowledge model in information retrieval. The standard method in which the term vector is an indicator vector that reflects the occurrence of terms in the document/query is on the left. Here all unobserved terms are treated as zeros. In contrast, our approach uses the knowledge model to fill in the values of unobserved terms with their probabilities.

where $P(e_h|d_i)$ is calculated from the PHITS model using Equation 1 and α is a constant that scales $P(e_h|d_i)$ to a new probability space. Constant α can be defined in various ways. In our work we assume α is

$$\alpha = 1/\min_j P(e_j|d)$$

, where j ranges over all entities explicitly mentioned in the document d.

As shown in Figure 1 we expect our domain knowledge inferences to be applied before standard information retrieval methods are deployed. The row vector at top of the figure is an indicator-based term-vector. This term vector is either transformed with the help of the knowledge model (our model) or kept, such that all unobserved terms are treated as zeros (standard model). Similarly to query expansion, we use the knowledge model to expand the term vectors for all unobserved concepts with their probabilities inferred from the PHITS models. A variety of information retrieval methods (e.g. PLSI(Hofmann 1999), LDA(Wei & Croft 2006), BM25(Robertson *et al.* 1994)) can be then applied to these two vector-term options.

Evaluation

To demonstrate the benefit of our approach we evaluate it on a PubMed-Cancer corpus that consists of 6000 PubMed articles on 10 common cancer. We use three information retrieval methods: LDA(Blei, Ng, & Jordan 2003), PLSI(Hofmann 1999) and BM25(Robertson *et al.* 1994) and try them both with and without our knowledge-driven transformation. We use Lucene as our baseline.

PubMed Cancer Corpus

The PubMed-cancer corpus and consists of 6,000 articles, and includes both full documents and their abstracts. In our information retrieval experiments we use abstracts; full documents are only used to assess the relevance of documents identified by the IR system.

Knowledge model

The association network representing our knowledge model was built using abstracts of 4800 training documents (80% of the PubMed Cancer corpus). The remaining 1200 documents in the corpus were used for testing and evaluation purposes. Figure 2 shows a small portion of the network that was extracted from the training documents. The graph depicts a closely related group of three species $ERBB$, TGF, and $HER3$. The complete network consists of over 51,000 association links in between pairs of gene or protein species. The association network was then used to build the PHITS model supporting various probabilistic inferences.

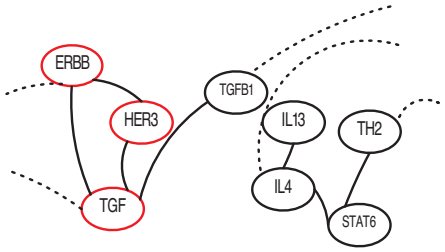


Figure 2: Snapshot of a part of the association network for the cancer document experiment

Evaluation method

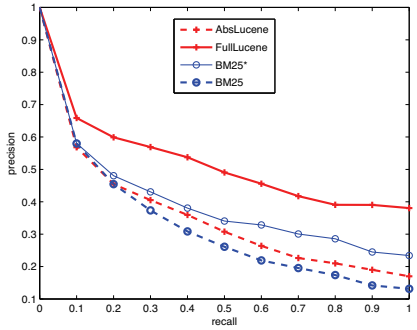
The relevance of a scientific document to the query, especially if partial matches need to be considered, is best assessed by a human expert. Unfortunately this is a very time-consuming process. To alleviate the problem and demonstrate the benefit of our approach in making useful inferences we adopt the following experimental 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, we retrieve a document based on its abstract, and the relevance of the document's abstract is judged (automatically) by the match of that query to the full document.

To evaluate the different IR methods we have generated a set of 500 queries that consisted of pairs of two species (proteins or genes) such that 100 of these queries were generated by randomly pairing any two species identified in the training corpus, and 400 queries were generated using documents in the testing corpus by the following process. To build a query we first randomly picked a testing document, and then randomly selected a pair of species that were associated with each other in the full text of this document. This helped us to generate queries that had a perfect match in the full text of at least one document. All 500 queries were run on abstracts only, the correctness of the retrieved document to the query was determined by analyzing the full text and the match of the query on the full text level.

Results

The queries were fed into three information-retrieval methods: BM25, PLSI and LDA. We have tried them both with and without our knowledge-model transformation. In addition, we compared the results to the Lucene engine (lucene.apache.org) which we use as a baseline. We run Lucene twice: first on abstracts, and then on full articles. AbsLucene runs on abstract and uses the same information as our methods and hence it forms the state-of-the-art baseline. FullLucene relies on the information in the full text that is not provided to other methods, hence it is expected to yield an upper bound baseline.

Figures 3- 5 show the interpolated precision-recall curves for the methods on 500 random queries. We use an asterisk ('*') to denote information-retrieval approaches enhanced by our 'knowledge-driven' transformation. As, expected the Lucene engine indexed on the full text performs the best, because it is the only system with access to full articles. We can see that the original BM25, PLSI and LDA do not outperform the AbsLucene baseline. This is not surprising because they use only domain concepts and deal with much more sparse matrix. A more interesting result is that all these approaches when they are enhanced by our method perform better than the baseline by a significant margin. Recall that we use the exact match in full text of a query to judge the relevance of abstracts. Although all these three techniques are not fully optimized, they still can outperform the well know search engine, Lucene. This result shows that the model can predict domain concepts that are expected to be mentioned in the full text from the knowledge extracted from the abstracts. We do not necessarily observe the occurrence of these domain concepts in the abstracts all the time, but the domain knowledge model helps us assess the chance of seeing them in the full text. This experiment shows that domain knowledge helps to improve the retrieval and supports our hypothesis that relevance is (at least partly) determined by tight connectivity of knowledge concepts. Moreover, it confirms that the domain knowledge is helpful to find relevant domain concepts.

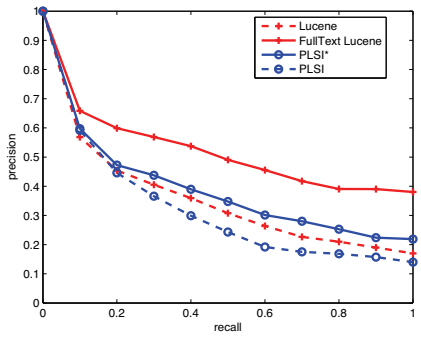


(a) BM25

Figure 3: Precision/Recall curves for the BM25 based methods. AbsLucene and FullLucene are used as baselines.

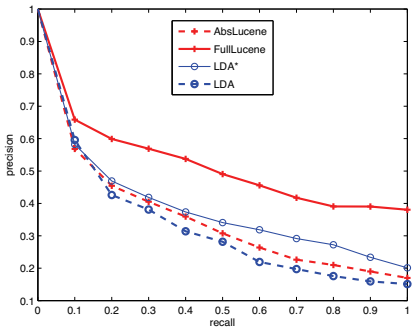
Related work

Information retrieval methods enhanced with domain knowledge were tried and used in multiple studies. For ex-



(a) PLSI

Figure 4: Precision/Recall curves for the PLSI-based information retrieval methods.



(a) LDA

Figure 5: Precision/Recall curves for the LDA-based information retrieval methods

ample, in (Pickens & MacFarlane 2006) the authors showed that the document terms can be weighted better with the help of context knowledge. (Zhou *et al.* 2007) studied various ways of incorporating domain knowledge from MeSH and Entrez Gene tools into the information retrieval process and demonstrated their improvement over baseline methods. (Bttcher, Clarke, & Cormack 2004) used the existing domain knowledge to expand the search queries. More specifically, the authors extracted synonyms of all biomedical terms from external databases and expanded the original queries using these terms. A related work by (Aronson & Rindflesch 1997) mapped the queries into biomedical concepts using MetaMap tools and added these concepts into the original queries. Finally, (Lin & Demner-Fushman 2006) showed the benefits of knowledge-base methods in the clinical medicine retrieval applications.

We note that all of the above approaches are different from our approach. We mine the domain knowledge from the document collection automatically and exploit it to infer the missing knowledge in individual documents.

Conclusion and Future Work

We have presented a new framework that extracts the domain knowledge from multiple documents and uses it to support document retrieval inferences. We showed that our method can improve the retrieval performance of documents in the biomedical literature. To our best knowledge this is the first work that attempts to learn the probabilistic relations among domain concepts via link analysis methods and apply them in the document retrieval.

The inference potential of our framework in retrieval of relevant documents was demonstrated on a surrogate experiment with document abstracts, in which full documents and relations therein were used only to assess quantitatively the relevance of the document to the query. Our approach can be combined with many existing techniques to improve document retrieval. We have shown the relative improvement on the PubMed dataset.

Our knowledge layer was extracted using associations among domain-specific terms observed in the document collection. We did not make any attempt to refine these associations and identify the relations they represent. However, we believe a more comprehensive domain knowledge with a variety of explicitly represented relations among the domain concepts and their analysis may further improve the information retrieval performance. Finally, a growing interest and work on the design and construction of domain ontology, opens up new possibilities for injecting the knowledge layer into information retrieval. Given the simplicity of association network, it is possible and easy to combine it with existing ontologies or databases such as KEGG pathway databases. Our model is robust enough to integrate knowledge from various sources and combine with existing retrieval methods.

Acknowledgement

This research was supported in part by grants R21 LM009102- 01A1 from NLM, P50 CA090440-06 from NCI and USAMRAA W81XWH-05-2-0066 from the Department of Defense.

References

Aronson, A. R., and Rindflesch, T. C. 1997. Query expansion using the umls metathesaurus. In *TREC '04: Proceedings of the AMIA Annual Fall Symposium 97, JAMIA Suppl*, 485–489. AMIA.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Bttcher, S.; Clarke, C. L. A.; and Cormack, G. V. 2004. Domain-specific synonym expansion and validation for biomedical information retrieval. In *TREC '04: Proceedings of the 13th Text REtrieval Conference*.

Cohn, D., and Chang, H. 2000. Learning to probabilistically identify authoritative documents. In *Proc. 17th International Conf. on Machine Learning*, 167–174. Morgan Kaufmann, San Francisco, CA.

Collins, M. 1999. Head-driven statistical models for natural language parsing. *PhD Dissertation*.

Consortium, T. G. O. 2000. Gene ontology: tool for the unification of biology. *Nature Genet* 25:25–29.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, 50–57. ACM Press.

Lin, J., and Demner-Fushman, D. 2006. The role of knowledge in conceptual retrieval: a study in the domain of clinical medicine. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 99–106. ACM.

Pickens, J., and MacFarlane, A. 2006. Term context models for information retrieval. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, 559–566. ACM.

Pillet, V.; Zehnder, M.; Seewald, A. K.; Veuthey, A.-L.; and Petra, J. 2005. Gpsdb: a new database for synonyms expansion of gene and protein names. *Bioinformatics* 21(8):1743–1744.

Robertson, S. E.; Walker, S.; Jones, S.; Hancock-Beaulieu, M.; and Gatford, M. 1994. Okapi at trec-3. In *In Proceedings of the Third Text REtrieval Conference (TREC 1994)*. Novemember.

Wei, X., and Croft, W. B. 2006. Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, 178–185. ACM.

Zhou, W.; Yu, C.; Smalheiser, N.; Torvik, V.; and Hong, J. 2007. Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval*, 655–662. ACM.