MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity

Angelos Hliaoutakis[1], Giannis Varelas[1], Euripides G.M. Petrakis[1], and Evangelos Milios[2]

[1] Dept. of Electronic and Computer Engineering
Technical University of Crete (TUC)
Chania, Crete, GR-73100, Greece
angelos@softnet.tuc.gr, varelas@softnet.tuc.gr,
petrakis@intelligence.tuc.gr
[2] Faculty of Computer Science, Dalhousie University
Halifax, Nova Scotia
B3H 1W5, Canada
eem@cs.dal.ca

Abstract. *MedSearch*[1] is a complete retrieval system for Medline, the premier bibliographic database of the U.S. National Library of Medicine (NLM). *MedSearch* implements *SSRM*, a novel information retrieval method for discovering similarities between documents containing semantically similar but not necessarily lexically similar terms.

1 Introduction

MedSearch is a complete retrieval system for medical literature. It supports retrieval by *SSRM* (*Semantic Similarity Retrieval Model*) [1], a novel information retrieval method which is capable for associating documents containing semantically similar (but not necessarily lexically similar) terms. *SSRM* suggests discovering semantically similar terms in documents and queries using term taxonomies (ontologies) and by associating such terms using semantic similarity methods (e.g., [2]). *SSRM* demonstrated very promising performance achieving significantly better precision and recall than Vector Space Model (VSM) for retrievals on Medline.

2 Semantic Similarity Retrieval Model (*SSRM*)

As it is typical in information retrieval, documents are represented by term vectors and each term is initially represented by its $tf \cdot idf$ weight. For short queries specifying only a few terms the weights are initialized to 1. Then, *SSRM* works in three steps:

[1] http://www.intelligence.tuc.gr/medsearch

Term Re-Weighting: The weight q_i of each query term i is adjusted based on its relationships with semantically similar terms j within the same vector

$$q_i = q_i + \sum_{\mathrm{sim}(i,j) \geq t}^{j \neq i} q_j \mathrm{sim}(i,j), \tag{1}$$

where t is a user defined threshold ($t = 0.8$ in this work). Semantic similarity between terms is computed according to the method described in [2]. Multiple related terms in the same query reinforce each other (e.g., "train", "metro"). The weights of non-similar terms remain unchanged (e.g., "train", "house").

Term Expansion: First, the query is augmented by synonym terms, using the most common sense of each query term. Then, the query is augmented by semantically similar terms higher or lower in the taxonomy (i.e., hypernyms and hyponyms). The neighborhood of the term in the taxonomy is examined and all terms with similarity greater than threshold T ($T = 0.9$ in this work) are also included in the query vector. This expansion may include terms more than one level higher or lower than the original term. Then, each query term i is assigned a weight as follows

$$q_i' = q_i + \sum_{\mathrm{sim}(i,j) \geq T \text{ and } j \in Q}^{i \neq j} \frac{1}{n} q_j \mathrm{sim}(i,j), \tag{2}$$

where n is the number of hyponyms of each expanded term j, q_i is the weight of term i before expansion and Q is the subset of the set of original query terms that led into new terms added to the expanded query. For hypernyms $n = 1$. Notice that $q_i = 0$ if term i was not in the original query but was introduced during the query expansion process. It is possible for a term to introduce terms that already existed in the query. It is also possible that the same term is introduced by more than one other terms. Eq. 2 suggests taking the weights of the original query terms into account and that the contribution of each term in assigning weights to query terms is normalized by the number n of its hyponyms. After expansion and re-weighting, the query vector is normalized by document length, like each document vector.

Document Similarity: The similarity between an expanded and re-weighted query q and a document d is computed as

$$Sim(q,d) = \frac{\sum_i \sum_j q_i d_j \mathrm{sim}(i,j)}{\sum_i \sum_j q_i d_j}, \tag{3}$$

where i and j are terms in the query and the document respectively. The similarity measure above is normalized in the range [0,1]. Notice that, if there are no semantically similar terms ($\mathrm{sim}(i,j) = 0 \ \forall i \neq j$) *SSRM* is reduced to VSM.

Expanding and re-weighting is fast for queries, which are typically short, consisting of only a few terms, but not for documents with many terms. The method suggests expansion of the query only. Notice that expansion with low threshold values T (e.g., $T = 0.5$) is likely to introduce many new terms and diffuse the topic of the query (topic drift).

3 *MedSearch*

Medline[2] is the premier bibliographic database of the U.S. National Library of Medicine (NLM), indexing more that 15 million references (version 2006) to journal articles in life sciences, medicine and bio-medicine. In addition to title, abstract and authors, Medline stores a rich set of metadata associated with each article such as language of publication, publication type, dates, source of publication and relations between articles. Articles in Medline are also indexed (manually by experts) by a set of descriptive MeSH terms.

MedSearch supports retrieval of bibliographic information on Medline by VSM as well as by semantic retrieval by *SSRM* using MeSH[3] as the underlying reference ontology. VSM and *SSRM* are implemented on top of Lucene[4] a full-featured text search engine library in Java. All documents are indexed by title, abstract and MeSH terms. These descriptions are syntactically analyzed and reduced into separate vectors of MeSH terms which are matched against the queries according to Eq. 3 (as similarity between expanded and re-weighted vectors). The weights of all MeSH terms are initialized to 1 while the weights of titles and abstracts are initialized by $tf \cdot idf$. The similarity between a query and a document is computed as

$$Sim(q, d) = Sim(q, d_{MeSH-terms}) + Sim(q, d_{title}) + Sim(q, d_{abstract}), \qquad (4)$$

where $d_{MeSH-terms}$, d_{title} and $d_{abstract}$ are the representations of the document MeSH terms, title and abstract respectively. This formula suggests that a document is similar to a query if most of its components are similar to the query.

The specification of threshold T in *SSRM* may depend on query scope or user uncertainty. A low value of T is desirable for broad scope queries or for initially resolving uncertainty as to what the user is really looking for. The query is then repeated with higher threshold. A high value of T is desirable for very specific queries: Users with high degree of certainty might prefer to expand with a high threshold or not to expand at all. The option is also adjustable at the interface of *MedSearch*. In this work we set $T = 0.9$ (i.e. the query is expanded only with very similar terms).

A set of 15 of medical queries were prepared by an independent medical expert. Each query specified between 3 and 10 terms and retrieved the best 20 answers. The results were evaluated by the same medical expert. Each method is represented by a *precision/recall* curve. Each point on a curve is the average precision and recall over all queries. Fig. 1 indicates that VSM with query expansion is obviously the worst method. Each query term is augmented by its "Entry Terms" in MeSH (i.e., general related terms which are not always synonyms). Notice that no exact synonymy relation is defined in MeSH. For this reason, in *SSRM* we do not apply expansion with Entry Terms. However, query terms

[2] http://www.ncbi.nlm.nih.gov/entrez

[3] http://www.nlm.nih.gov/mesh

[4] http://lucene.apache.org

are expanded with semantically similar terms in the neighborhood of each term according to Eq. 2. Semantic information retrieval by *SSRM* is more effective than classic information retrieval by VSM achieving up to 20% better precision and up to 20% better recall.

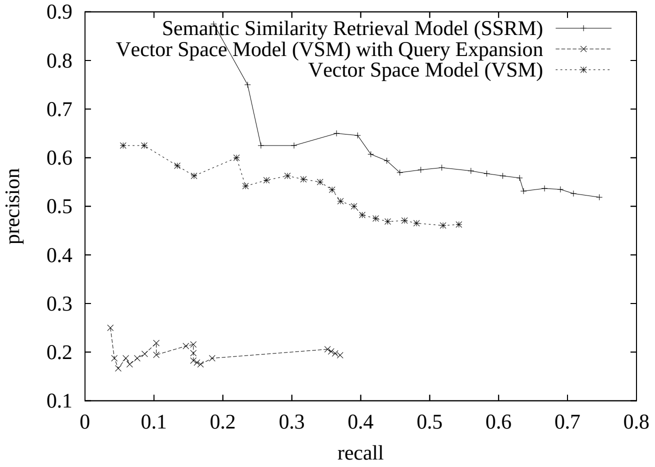

Fig. 1. Precision-recall diagram of *SSRM* and VSM for retrievals on Medline.

4 Conclusions

MedSearch is an information retrieval system for medical literature and is accessible on the Web. *MedSearch* supports retrieval by VSM (the classic retrieval method) and by *SSRM*. *SSRM* demonstrates promising performance improvements over VSM. *SSRM* can work in conjunction with any taxonomic ontology (e.g. WordNet).

Acknowledgement

We thank Dr. Qi, Dalhousie University, for her help in preparing the set of medical queries. This work was supported by project BIOPATTERN (FP6-NoE, Project No. 508803) of the European Union (EU).

References

1. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., Milios, E.: Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In: 7^{th} ACM Intern. Workshop on Web Information and Data Management (WIDM 2005), Bremen, Germany (2005) 10–16
2. Leacock, C., Chodorow, M.: Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet. In Fellbaum, C., ed.: An Electronic Lexical Database. MIT Press (1998) 265–283