Information Retrieval by Semantic Similarity

Angelos Hliaoutakis[1], Giannis Varelas[1], Epimeneidis Voutsakis[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. Semantic Similarity relates to computing the similarity between conceptually similar but not necessarily lexically similar terms. Typically, semantic similarity is computed by mapping terms to an ontology and by examining their relationships in that ontology. We investigate approaches to computing the semantic similarity between natural language terms (using WordNet as the underlying reference ontology) and between medical terms (using the MeSH ontology of medical and biomedical terms). The most popular semantic similarity methods are implemented and evaluated using WordNet and MeSH. Building upon semantic similarity we propose the Semantic Similarity based Retrieval Model (*SSRM*), a novel information retrieval method capable for discovering similarities between documents containing conceptually similar terms. The most effective semantic similarity method is implemented into *SSRM*. *SSRM* has been applied in retrieval on OHSUMED (a standard TREC collection available on the Web). The experimental results demonstrated promising performance improvements over classic information retrieval methods utilizing plain lexical matching (e.g., Vector Space Model) and also over state-of-the-art semantic similarity retrieval methods utilizing ontologies.

key-words: Information Retrieval, Semantic Similarity, WordNet, MeSH, Ontology

1 Introduction

Semantic Similarity relates to computing the similarity between concepts which are not necessarily lexically similar. Semantic similarity aims at providing robust tools for standardizing the

⋆ Corresponding author.

content and delivery of information across communicating information sources. This has long been recognized as a central problem in Semantic Web where related sources need to be linked and communicate information to each other. Semantic Web will also enable users to retrieve information in a more natural and intuitive way (as in a "query-answering" interaction).

In the existing Web, information is acquired from several disparate sources in several formats (mostly text) using different language terminologies. Interpreting the meaning of this information is left to the users. This task can be highly subjective and time consuming. To relate concepts or entities between different sources (the same as for answering user queries involving such concepts or entities), the concepts extracted from each source must be compared in terms of their meaning (i.e. semantically). Semantic similarity offer the means by which this this goal can be realized.

This work deals with a certain aspect of Semantic Web and semantics, that of semantic text association and text semantics respectively. We demonstrate that it is possible to approximate algorithmically the human notion of similarity using semantic similarity and to develop methods capable of detecting similarities between conceptually similar documents even when they don't contain lexically similar terms. The lack of common terms in two documents does not necessarily mean that the documents are not related. Computing text similarity by classical information retrieval models (e.g., Vector Space, Probabilistic, Boolean) [1] is based on lexical term matching. However, two terms can be semantically similar (e.g., can be synonyms or have similar meaning) although they are lexically different. Therefore, classical retrieval methods will fail to associate documents with semantically similar but lexically different terms.

In the context of the multimedia semantic web, our method would permit informal textual descriptions of multimedia content to be effectively used in retrieval, and obviates the need for generating structured metadata. Informal descriptions require significantly less human labor than structured descriptions.

In the first part of this work we present a critical evaluation of several semantic similarity approaches for computing the semantic similarity between terms using two well known taxonomic

hierarchies namely WordNet [3] and MeSH[4]. WordNet is a controlled vocabulary and thesaurus offering a taxonomic hierarchy of natural language terms developed at Princeton University. MeSH (Medical Subject Heading) is a controlled vocabulary and a thesaurus developed by the U.S. National Library of Medicine (NLM)[5] offering a hierarchical categorization of medical terms. Similar results for MeSH haven't been reported before in the literature. All methods are implemented and integrated into a semantic similarity system which is accessible on the Web[6].

In the second part of this work we propose the *Semantic Similarity Retrieval Model* (*SSRM*). *SSRM* suggests discovering semantically similar terms in documents (e.g., between documents and queries) using general or application specific term taxonomies (like WordNet or MeSH) and by associating such terms using semantic similarity methods. Initially, *SSRM* computes $tf \cdot idf$ weights to term representations of documents. These representations are then augmented by semantically similar terms (which are discovered from WordNet or MeSH by applying a range query in the neighborhood of each term in the taxonomy) and by re-computing weights to all new and pre-existing terms. Finally, document similarity is computed by associating semantically similar terms in the documents and in the queries respectively and by accumulating their similarities.

SSRM together with the term-based Vector Space Model (VSM) [2] (the classic document retrieval method utilizing plain lexical similarity) as well as the most popular semantic information retrieval methods in the literature (i.e., [2–4]) are all implemented and evaluated on OHSUMED [5] (a standard TREC collection with 293,856 medical articles) and on a crawl of the Web with more than 1.5 million Web pages with images. *SSRM* demonstrated very promising performance achieving significantly better precision and recall than its competitors.

The rest of this paper in organized as follows: A review of related work on semantic information retrieval is presented in Sec. 2. WordNet and MeSH, the two taxonomic hierarchies used in this work, as well as a critical evaluation of several semantic similarity methods on WordNet

[3] http://wordnet.princeton.edu

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

[5] http://www.nlm.nih.gov

[6] http://www.intelligence.tuc.gr/similarity

and MeSH is presented in Sec. 3. *SSRM* the proposed semantic similarity based retrieval model is presented in Sec. 4 followed by Conclusions and issues for further research in Sec. 5.

2 Related Work

Query expansion with potentially related (e.g., similar) terms has long been considered a means for resolving term ambiguities and for revealing the hidden meaning in user queries. A recent contribution by Collins-Thomson [6] proposed a framework for combining multiple knowledge sources for revealing term associations and for determining promising terms for query expansion. Given a query, a term network is constructed representing the relationships between query and potentially related terms obtained by multiple knowledge sources such as synonym dictionaries, general word association scores, co-occurrence relationships in corpus or in retrieved documents. In the case of query expansion, the source terms are the query terms and the target terms are potential expansion terms connected with the query terms by labels representing probabilities of relevance. The likelihood of relevance between such terms is computed using random walks and by estimating the probability of the various aspects of the query that can be inferred from potential expansion terms. *SSRM* is complementary to this approach: It shows how to handle more relationship types (e.g., hyponyms, hypernyms in an ontology) and how to compute good relevance weights given the $tf \cdot idf$ weights of the initial query terms. *SSRM* focuses on semantic relationships (a specific aspect of term relationships not considered in [6]) and demonstrates that that it is possible to enhance the performance of retrievals using this information alone.

SSRM is also complementary to work by Voorhees [3] as well as to work by Richardson and Smeaton [4]. Voorhees [3] proposed expanding query terms with synonyms, hyponyms and hypernyms in WordNet but did not propose an analytic method for setting the weights of these terms. Voorhees reported some improvement for short queries, but little or no improvement for long queries. Richardson and Smeaton [4] proposed taking the summation of the semantic similarities between all possible combinations of document and query terms. They ignored the relevant significance of terms (as captured by $tf \cdot idf$ weights) and they considered neither term

expansion nor re-weighting. Our proposed method takes term weights into account, introduces an analytic and intuitive term expansion and re-weighting method and suggests a document similarity formula that takes the above information into account. Similarly to *SSRM*, the text retrieval method by Mihalcea [7] works by associating only the most semantically similar terms in two documents and by summing up their semantic similarities (weighted by the inverse document frequency idf). Query terms are not expanded nor re-weighted as in *SSRM*. Notice that *SSRM* associates all terms in the two documents and accumulates their semantic similarities.

The methods referred to above allow for ordering the retrieved documents by decreasing similarity to the query taking into account that two documents may match only partially (i.e., a retrieved document need not contain all query terms). Similarly to classic retrieval models like VSM, *SSRM* allows for non-binary weights in queries and in documents (initial weights are computed using the standard $tf \cdot idf$ formula). The experimental results in this work demonstrate that *SSRM* performs significantly better (achieving better precision and recall) than its competitors (i.e., VSM [2] and ontology-based methods [2–4]).

Query expansion and term re-weighting in *SSRM* resemble also earlier approaches which attempt to improve the query with terms obtained from a similarity thesaurus (e.g., based on term to term relationships [8, 9]). This thesaurus is usually computed by automatic or semi-automatic corpus analysis (global analysis) and would not only add new terms to *SSRM* but also reveal new relationships not existing in a taxonomy of terms. This approach depends on the corpus.

SSRM is independent of the corpus and works by discovering term associations based on their conceptual similarity in a lexical ontology specific to the application domain at hand (i.e., WordNet or MeSH in this work). The proposed query expansion scheme is complementary to methods which expand the query with co-occurrent terms (e.g., “railway”, “station”) in retrieved documents [10] (local analysis). Expansion with co-occurrent terms (the same as a thesaurus like expansion) can be introduced as additional expansion step in the method. Along the same lines, *SSRM* needs to be extended to work with phrases [11]. Along the same lines, the method by Possas et.al. [12] exploits the intuition that co-occurrent terms occur close to each other

and propose a method for extracting patterns of co-occurrent terms and their weights by data mining.

3 Semantic Similarity

Issues related to semantic similarity algorithms along with issues related to computing semantic similarity on WordNet and MeSH are discussed below.

3.1 WordNet

WordNet [7] is an on-line lexical reference system developed at Princeton University. WordNet attempts to model the lexical knowledge of a native speaker of English. WordNet can also be seen as an ontology for natural language terms. It contains around 100,000 terms, organized into taxonomic hierarchies. Nouns, verbs, adjectives and adverbs are grouped into synonym sets (synsets). The synsets are also organized into senses (i.e., corresponding to different meanings of the same term or concept). The synsets (or concepts) are related to other synsets higher or lower in the hierarchy defined by different types of relationships. The most common relationships are the *Hyponym/Hypernym* (i.e., Is-A relationships), and the *Meronym/Holonym* (i.e., Part-Of relationships). There are nine noun and several verb Is-A hierarchies (adjectives and adverbs are not organized into Is-A hierarchies). Fig. 1 illustrates a fragment of the WordNet Is-A hierarchy.

3.2 MeSH

MeSH[8] (Medical Subject Headings) is a taxonomic hierarchy (ontology) of medical and biological terms (or concepts) suggested by the U.S National Library of Medicine (NLM). MeSH terms are organized in Is-A taxonomies with more general terms (e.g "chemicals and drugs") higher in a taxonomy than more specific terms (e.g "aspirin"). There are 15 taxonomies with more than 22,000 terms. A term may appear in more than one taxonomy. Each MeSH term is

[7] http://wordnet.princeton.edu

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

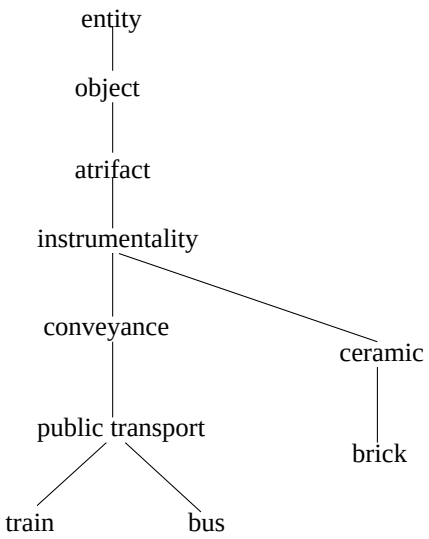

Fig. 1. A fragment of the WordNet Is-A hierarchy.

described by several properties, the most important of them being the *MeSH Heading (MH)* (i.e., term name or identifier), *Scope Note* (i.e., a text description of the term) and *Entry Terms* (i.e., mostly synonym terms to the MH). Entry terms also include stemmed MH terms and are sometimes referred to as quasi-synonyms (they are not always exactly synonyms). Each MeSH terms is also characterized by its MeSH tree number (or code name) indicating the exact position of the term in the MeSH tree taxonomy (e.g., "D01,029" is the code name of term "Chemical and drugs"). Fig. 2 illustrates a fragment of the MeSH Is-A hierarchy.

3.3 Semantic Similarity Methods

Several methods for determining semantic similarity between terms have been proposed in the literature and most of them have been tested on WordNet [9]. Similar results on MeSH haven't been reported in the literature.

Semantic similarity methods are classified into four main categories:

Edge Counting Methods: Measure the similarity between two terms (concepts) as a function of the length of the path linking the terms and on the position of the terms in the taxonomy [13–17].

[9] http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi

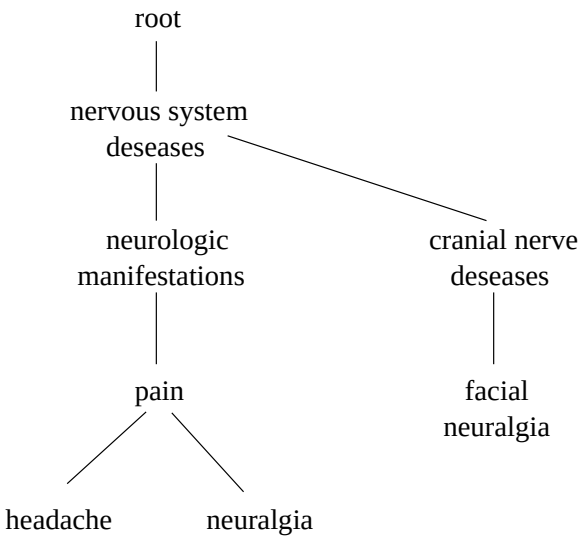

Fig. 2. A fragment of the MeSH Is-A hierarchy.

Information Content Methods: Measure the difference in information content of the two terms as a function of their probability of occurrence in a corpus [18–21]. In this work information content is computed according to [22]: MeSH is used as a statistical resource for computing the probabilities of occurrence of terms: More general concepts (higher in the hierarchy) with many hyponyms convey less information content than more specific terms (lower in the hierarchy) with less hyponyms. This approach is independent of the corpus and also guarantees that the information content of each term is less than the information content of its subsumed terms. This constraint is common to all methods of this category. Computing information content from a corpus does not always guarantee this requirement. The same method is also applied for computing the information content of MeSH terms.

Feature based Methods: Measure the similarity between two terms as a function of their properties (e.g., their definitions or "glosses" in WordNet or "scope notes" in MeSH) or based on their relationships to other similar terms in the taxonomy. Common features tend to increase the similarity and (conversely) non-common features tend to diminish the similarity of two concepts [23].

Hybrid methods combine the above ideas [24]: Term similarity is computed by matching synonyms, term neighborhoods and term features. Term features are further distinguished into parts, functions and attributes and are matched similarly to [23].

Semantic similarity methods can also be distinguished between:

Single Ontology similarity methods, which assume that the terms which are compared are from the same ontology (e.g., MeSH).

Cross Ontology similarity methods, which compare terms from two different ontologies (e.g., WordNet and MeSH).

An important observation and a desirable property of most semantic similarity methods is that they assign higher similarity to terms which are close together (in terms of path length) and lower in the hierarchy (more specific terms), than to terms which are equally close together but higher in the hierarchy (more general terms).

Edge counting and information content methods work by exploiting structure information (i.e., position of terms) and information content of terms in a hierarchy and are best suited for comparing terms from the same ontology. Because the structure and information content of different ontologies are not directly comparable, cross ontology similarity methods usually call for hybrid or feature based approaches. The focus of this work is on single ontology methods. For details on the methods used in the work please refer to [25].

Additional properties of the similarity referred to above are summarized in Table 1. It shows, method type, whether similarity affected by the common characteristics of the concepts which are compared, whether it decreases with their differences, whether the similarity is a symmetric property, whether its value is normalized in $[0, 1]$ and, finally, whether it is affected by the position of the terms in the taxonomy.

3.4 Semantic Similarity System

All methods above are implemented and integrated into a semantic similarity system which is available on the Web [10]. Fig. 3 illustrates the architecture of this system. The system communi-

[10] http://www.intelligence.tuc.gr/similarity

Method	Method Type	Increases with Commonality	Decreases with Difference	Symmetric Property	Normalized in $[0, 1]$	Position in hierarchy
Rada [13]	Edge Counting	yes	yes	yes	yes	no
Wu [14]	Edge Counting	yes	yes	yes	yes	yes
Li [15]	Edge Counting	yes	yes	yes	yes	yes
Leacock [16]	Edge Counting	no	yes	yes	no	yes
Richardson [17]	Edge Counting	yes	yes	yes	yes	yes
Resnik [19]	Info. Content	yes	no	yes	no	yes
Lin [20]	Info. Content	yes	yes	yes	yes	yes
Lord [18]	Info. Content	yes	no	yes	yes	yes
Jiang [21]	Info. Content	yes	yes	yes	no	yes
Tversky [23]	Feature	yes	yes	no	yes	no
Rodriguez [24]	Hybrid	yes	yes	no	yes	no

Table 1. Summary of semantic similarity methods.

cates with WordNet and MeSH. Each term is represented by its tree hierarchy (corresponding to an XML file) which is stored in the XML repository. The tree hierarchy of a term represents the relationships of the term with its hyponyms and hypernyms. These XML files are created by the XML generator using the WordNet XML Web-Service[11]. The purpose of this structure is to facilitate access to terms stored in the XML repository by indexing the terms by their name of identifier (otherwise accessing a term would require exhaustive searching through the entire WordNet or MeSH files). The information content of all terms is also computed in advance and stored separately in the information content database. The user is provided with several options at the user interface (e.g., sense selection, method selection).

3.5 Evaluation of Semantic Similarity Methods

In the following we present a comparative evaluation of the similarity methods referred to in Sec. 3.3.

[11] http://wnws.sourceforge.net

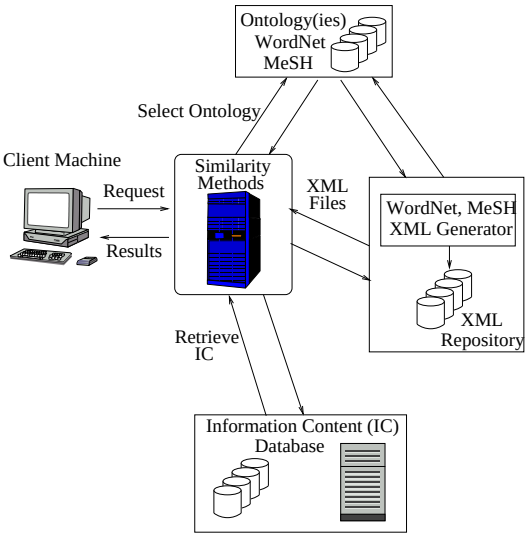

Fig. 3. Semantic Similarity System.

Semantic Similarity on WordNet: In accordance with previous research [19], we evaluated the results obtained by applying the semantic similarity methods of Sec. 3.3 to the same pairs used in the experiment by Miller and Charles [26]: 38 undergraduate students were given 30 pairs of nouns and were asked to rate the similarity of each pair on a scale from 0 (not similar) through 4 (perfect synonymy). The average rating of each pair represents a good estimate of how similar the two words are.

We compared the computed similarity scores for the same terms as in Miller and Charles with the human relevance results reported there. The similarity values obtained by all competitive computational methods (all senses of the first term are compared with all senses of the second term) are correlated with the average scores obtained by the humans in [26]. The higher the correlation of a method the better the method is (i.e., the more it approaches the results of human judgements).

Table 2 shows the correlation obtained by each method. Jiang and Conrath [21] suggested removing one of the pairs from the evaluation. This increased the correlation of their method to 0.87. The method by Li et. al. [15] is among the best and it is also the fastest. These results lead to the following observations:

Method	Method Type	Correlation
Rada [13]	Edge Counting	0.59
Wu [14]	Edge Counting	0.74
Li [15]	Edge Counting	0.82
Leacock [16]	Edge Counting	0.82
Richardson [17]	Edge Counting	0.63
Resnik [19]	Information Content	0.79
Lin [20]	Information Content	0.82
Lord [18]	Information Content	0.79
Jiang [21]	Information Content	0.83
Tversky [23]	Feature	0.73
Rodriguez [24]	Hybrid	0.71

Table 2. Evaluation of Edge Counting, Information Content, Feature based and Hybrid semantic similarity methods on WordNet.

- Information Content methods perform very well and close to the upper bound suggested by Resnik [19].
- Methods that consider the positions of the terms in the hierarchy (e.g., [15]), perform better than plain path length methods(e.g., [13]).
- Methods exploiting the properties (i.e., structure and information content) of the underlying hierarchy perform better than Hybrid and Feature based methods, which do not fully exploit this information. However, Hybrid and feature based methods (e.g., [24]) are mainly targeted towards cross ontology similarity applications where edge counting and information content methods do not apply.

Semantic Similarity on MeSH: An evaluation of Semantic Similarity methods on MeSH haven't been reported in the literature before. For the evaluation, we designed an experiment similar to that by Miller and Charles [26] for WordNet: We asked a medical expert to compile a set of MeSH term pairs. A set of 49 pairs was proposed, together with an estimate of similarity between 0 (not similar) and 4 (perfect similarity) for each pair. To reduce the subjectivity of

similarity estimates, we created a form-based interface with all pairs on the Web[12] and we invited other medical experts to enter their evaluation (the interface is still accepting results by experts world-wide). So far we received estimates from 12 experts.

The analysis of the results revealed that: (a) Some medical terms are more involved, or ambiguous leading to ambiguous evaluation by many users. For each pair, the standard deviation of their similarity (over all users) was computed. Pairs with standard deviation higher than a user defined threshold $t = 0.8$ were excluded from the evaluation. (b) Medical experts were not at the same level of expertise and (in some cases) gave unreliable results. For each user we computed the standard deviation of their evaluation (over all pairs). We excluded users who gave significantly different results from the majority of other users. Overall, 13 out of the 49 pairs and 4 out of the 12 users were excluded from the evaluation.

Following the same procedure as in the WordNet experiments, the similarity values obtained by each method (all senses of the first term are compared with all senses of the second term) are correlated with the average scores obtained by the humans. The correlation results are summarized in Table 3. These results lead to the following observations:

– Edge counting and information content methods perform about equally well. However, methods that consider the positions of the terms (lower or higher) in the hierarchy (e.g., [15]), perform better than plain path length methods(e.g., [13, 17]).

– Hybrid and feature based methods exploiting properties of terms (e.g., scope notes, entry terms) perform at least as well as information content and edge counting methods (exploiting information relating to the structure and information content of the underlying taxonomy), implying that term annotations in MeSH represent significant information by themselves and that it is possible to design even more effective methods by combining information from all the above sources (term annotations, structure information and information content).

[12] http://www.intelligence.tuc.gr/mesh

Method	Method Type	Correlation
Rada [13]	Edge Counting	0.50
Wu [14]	Edge Counting	0.67
Li [15]	Edge Counting	0.70
Leacock [16]	Edge Counting	0.74
Richardson [17]	Edge Counting	0.64
Resnik [19]	Information Content	0.71
Lin [20]	Information Content	0.72
Lord [18]	Information Content	0.70
Jiang [21]	Information Content	0.71
Tversky [23]	Feature	0.67
Rodriguez [24]	Hybrid	0.71

Table 3. Evaluation of Edge Counting, Information Content, Feature based and Hybrid semantic similarity methods on MeSH.

4 Semantic Similarity Retrieval Model (*SSRM*)

Traditionally, the similarity between two documents (e.g., a query q and a document d) is computed according to the Vector Space Model (VSM) [2] as the cosine of the inner product between their document vectors

$$Sim(q,d) = \frac{\sum_i q_i d_i}{\sqrt{\sum_i q_i^2}\sqrt{\sum_i d_i^2}}, \quad (1)$$

where q_i and d_i are the weights in the two vector representations. Given a query, all documents are ranked according to their similarity with the query. This model is also known as the *bag of words* model for document retrieval.

The lack of common terms in two documents does not necessarily mean that the documents are unrelated. Semantically similar concepts may be expressed in different words in the documents and the queries, and direct comparison by word-based VSM is not effective. For example, VSM will not recognize synonyms or semantically similar terms (e.g., “car”, “automobile”).

SSRM suggests discovering semantically similar terms using term taxonomies like WordNet or MeSH. Query expansion is also applied as a means for capturing similarities between terms of different degrees of generality in documents and queries (e.g., “human”, “man”). Queries are augmented with conceptually similar terms which are retrieved by applying a range query

in the neighborhood of each term in an ontology. Each query term is expanded by synonyms, hyponyms and hypernyms. The degree of expansion is controlled by the user (i.e., so that each query term may introduce new terms more than one level higher or lower in an ontology). *SSRM* can work with any general or application specific ontology. The selection of ontology depends on the application domain (e.g., WordNet for image retrieval on the Web [27], MeSH for retrieval in medical document collections [28]).

Query expansion by *SSRM* resembles the idea by Voorhees [3]. However, Voorhees did not show how to compute good weights for the new terms introduced into the query after expansion nor it showed how to control the degree of expansion (i.e., degree of expansion results in topic drift). *SSRM* solves exactly this problem. *SSRM* implements an intuitive and analytic method for setting the weights of the new query terms. Voorhees relied on the Vector Space Model (VSM) and therefore on lexical term matching for computing document similarity. Therefore, it is not possible for this method to retrieve documents with conceptually similar but lexically different terms. *SSRM* solves this problem by taking all possible term associations between two documents into account and by accumulating their similarities.

Similarly to VSM, queries and documents are first syntactically analyzed and reduced into term vectors. Very infrequent or very frequent terms are eliminated. Each term in this vector is represented by its weight. The weight of a term is computed as a function of its frequency of occurrence in the document collection and can be defined in many different ways. The term frequency - inverse document frequency ($tf \cdot idf$) model [2] is used for computing the weight. Typically, the weight d_i of a term i in a document is computed as

$$d_i = tf_i \cdot idf_i, \tag{2}$$

where tf_i is the frequency of term i in the document and idf_i is the inverse document frequency of i in the whole document collection. The formula is slightly modified for queries to give more emphasis to query terms.

Then *SSRM* works in three steps:

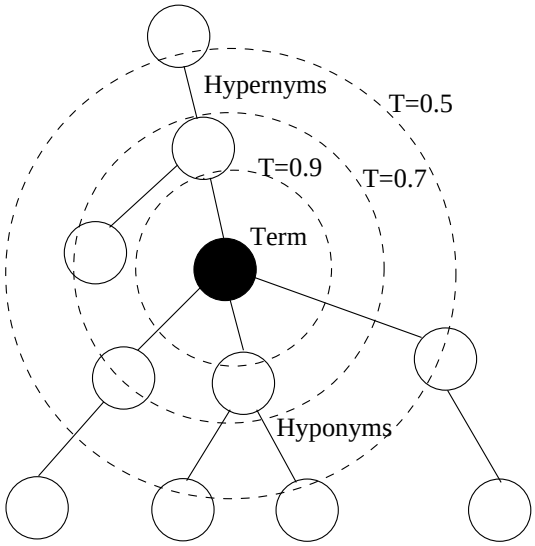

Fig. 4. Term expansion.

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

$$q_i' = q_i + \sum_{\text{sim}(i,j) \geq t}^{j \neq i} q_j \text{sim}(i,j), \tag{3}$$

where t is a user defined threshold ($t = 0.8$ in this work). Multiple related terms in the same query reinforce each other (e.g., "railway", "train", "metro"). The weights of non-similar terms remain unchanged (e.g., "train", "house"). For short queries specifying only a few terms the weights are initialized to 1 and are adjusted according to the above formula.

Query Expansion: First, the query is augmented by synonym terms, using the most common sense of each query term. Then, the query is augmented by terms higher or lower in the tree hierarchy (i.e., hypernyms and hyponyms) which are semantically similar to terms already in the query. Fig. 4 illustrates this process: Each query term is represented by its tree hierarchy. The neighborhood of the term is examined and all terms with similarity greater than threshold T 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

$$q'_i = \begin{cases} \sum_{\text{sim}(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j \text{sim}(i,j), & i \text{ is a new term} \\ q_i + \sum_{\text{sim}(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j \text{sim}(i,j), & i \text{ had weight } q_i, \end{cases} \tag{4}$$

where n is the number of hyponyms of each expanded term j. For hypernyms $n = 1$. The summation is taken over all terms j introducing terms to the query. 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. 6 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

$$Sim(q,d) = \frac{\sum_i \sum_j q_i d_j \text{sim}(i,j)}{\sum_i \sum_j q_i d_j}, \tag{5}$$

where i and j are terms in the query and the document respectively. Query terms are expanded and re-weighted according to the previous steps while document terms d_j are computed as $tf \cdot idf$ terms (they are neither expanded nor re-weighted). The similarity measure above is normalized in the range [0,1].

Fig. 5 presents a summary of *SSRM*.

4.1 Discussion

SSRM relaxes the requirement of classical retrieval models that conceptually similar terms be mutually independent (known also as "synonymy problem"). It takes into account dependencies between terms during its expansion and re-weighting steps. Their dependence is expressed quantitatively by virtue of their semantic similarity and this information is taken explicitly into account in the computation of document similarity. Notice however the quadratic time complexity of *SSRM* due to Eq. 5 as opposed to the linear time complexity of Eq. 1 of VSM. To speed

Input: Query q, Document d, Semantic Similarity sim, Thresholds t, T, Ontology.

Output: Document similarity value $Sim(d, q)$.

1. Compute Query term vector: $(q_1, q_2, \ldots)$ using $tf \cdot idf$ weighting scheme.

2. Compute Document term vector: $(d_1, d_2, \ldots)$ using $tf \cdot idf$ weighting scheme.

3. Query Re-weighting: For all terms i in query compute new weight $q'_i = q_i + \sum_{\text{sim}(i,j)\geq t}^{j\neq i} q_j \text{sim}(i, j)$.

4. Query Expansion: For all terms j in query retrieve terms i from ontology satisfying $sim(i, j) \geq T$.

5. Term Weighting: For all terms i in query compute new weight as

$$q'_i = \begin{cases} \sum_{\text{sim}(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j \text{sim}(i, j), & i \text{ is a new term} \\ q_i + \sum_{\text{sim}(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j \text{sim}(i, j), & i \text{ had weight } q_i, \end{cases} \quad (6)$$

6. Query Normalization: Normalize query by query length.

7. Compute Document Similarity: $Sim(q, d) = \frac{\sum_i \sum_j q_i d_j \text{sim}(i,j)}{\sum_i \sum_j q_i d_j}$.

Fig. 5. *SSRM* Algorithm.

up similarity computations, the semantic similarities between pairs of MeSH or WordNet terms are stored in a hash table. To reduce space only pairs with similarity greater than 0.3 are stored.

SSRM approximates VSM in the case of non-semantically similar terms: If $\text{sim}(i, j) = 0 \; \forall i \neq j$ then Eq. 5 is reduced to Eq. 1. In this case, the similarity between two documents is computed as a function of weight similarities between identical terms (as in 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. However, the similarity function will take into account the relationships between all semantically similar terms between the document and the query (something that VSM cannot do).

The expansion step attempts to automate the manual or semi-automatic query re-formulation process based on feedback information from the user [29]. Expanding the query with a threshold T will introduce new terms depending also on the position of the terms in the taxonomy: More specific terms (lower in the taxonomy) are more likely to expand than more general terms (higher in the taxonomy). 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). The spec-

ification of threshold T may also depend on query scope or user uncertainty. A low value of T might be 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. High values of threshold are 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 specification of T in Eq. 6 requires further investigation. Appropriate threshold values can be learned by training or relevance feedback [30]. Word sense disambiguation [31] can also be applied to detect the correct sense to expand rather than expanding the most common sense of each term. *SSRM* also makes use of a second threshold t for expressing the desired similarity between terms within the a query (Eq. 3 and Eq. 5). Our experiments with several values of t revealed that the method is rather insensitive to the selection of this threshold. Throughout this work we set $t = 0.8$.

4.2 Evaluation of *SSRM*

SSRM has been tested on two different applications and two data sets respectively. The first application is retrieval of medical documents using MeSH and the second application is image retrieval on the Web using WordNet. The experimental results below illustrate that it is possible to enhance the quality of classic information retrieval methods by incorporating semantic similarity within the retrieval method. *SSRM* outperforms classic and state-of-the-art semantic information retrieval methods [2–4]. The retrieval system is built upon Lucene[13], a full-featured text search engine library written in Java. All retrieval methods are implemented on top of Lucene.

The following methods are implemented and evaluated:

Semantic Similarity Retrieval Model (*SSRM*): Queries are expanded with semantically similar terms in the neighborhood of each term. The results below correspond to two different thresholds $T = 0.9$ (i.e. the query is expanded only with very similar terms) and $T = 0.5$ (i.e., the query is expanded with terms which are not necessarily conceptually similar). In

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

WordNet, each query term is also expanded with synonyms. Because no synonymy relation is defined in MeSH we did not apply expansion to Mesh terms in the query with Entry Terms. Semantic similarity in *SSRM* is computed by Li et.al. [15].

Vector Space Model (VSM) [2]: Text queries can also be augmented by synonyms.

Voorhees [3]: The query terms are expanded always with hyponyms one level higher or lower in the taxonomy and synonyms. The method did not propose an analytic method for computing the weights of these terms.

Richardson et.al. [4]: Accumulates the semantic similarities between all pairs document and query terms. It ignores the relative significance of terms (as it is captured by $tf \cdot idf$). Query terms are not expanded nor re-weighted as in *SSRM*.

In the experiments below, each method is represented by a *precision/recall* curve. For each query, the best 50 answers were retrieved (the precision/recall plot of each method contains exactly 50 points). Precision and recall values are computed from each answer set and therefore, each plot contains exactly 50 points. The top-left point of a precision/recall curve corresponds to the precision/recall values for the best answer or best match (which has rank 1), while the bottom right point corresponds to the precision/recall values for the entire answer set. A method is better than another if it achieves better precision and better recall. As we shall see in the experiments, it is possible for two precision-recall curves to cross-over. This means that one of the two methods performs better for small answer sets (containing less answers than the number of points up to the cross-section), while the other performs better for larger answer sets. The method achieving higher precision and recall for the first few answers is considered to be the better method (based on the assumption that typical users focus their attention on the first few answers).

Information Retrieval on OHSUMED: *SSRM* has been tested on OHSUMED[14] (a standard TREC collection with 293,856 medical articles from Medline published between 1988-1991) using MeSH as the underlaying ontology. All OHSUMED documents are indexed by title, abstract and MeSH terms (MeSH Headings). These descriptions are syntactically analyzed and

[14] http://trec.nist.gov/data/t9_filtering.html

reduced into separate vectors of MeSH terms which are matched against the queries according to Eq. 5 (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

$$Sim(q, d) = Sim(q, d_{MeSH-terms}) + Sim(q, d_{title}) + Sim(q, d_{abstract}), \quad (7)$$

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 its components are similar to the query. Each similarity component can be computed either by VSM or by *SSRM*.

For the evaluations, we applied the subset of 63 queries of the original query set developed by Hersh et al. [5]. The correct answers to these queries were compiled by the editors of OHSUMED and are also available from TREC[15] along with the queries.

The results in Fig. 6 demonstrate that *SSRM* with expansion with very similar terms ($T = 0.9$) and for small answer sets (i.e., with less than 8 answers) outperforms all other methods [2, 4, 3]. For larger answer sets, Voorhees [3] is the best method. For answer sets with 50 documents all methods (except VSM) perform about the same. *SSRM* with expansion threshold $T = 0.5$ performed worse than *SSRM* with $T = 0.9$. An explanation may be that it introduced many new terms and not all of them are conceptually similar with the original query terms.

Image Retrieval on the Web: Searching for effective methods to retrieve information from the Web has been in the center of many research efforts during the last few years. The relevant technology evolved rapidly thanks to advances in Web systems technology [32] and information retrieval research [33]. Image retrieval on the Web, in particular, is a very important problem in itself [34]. The relevant technology has also evolved significantly propelled by advances in image database research [35].

Image retrieval on the Web requires that content descriptions be extracted from Web pages and used to determine which Web pages contain images that satisfy the query selection crite-

[15] http://trec.nist.gov/data/t9_filtering.html

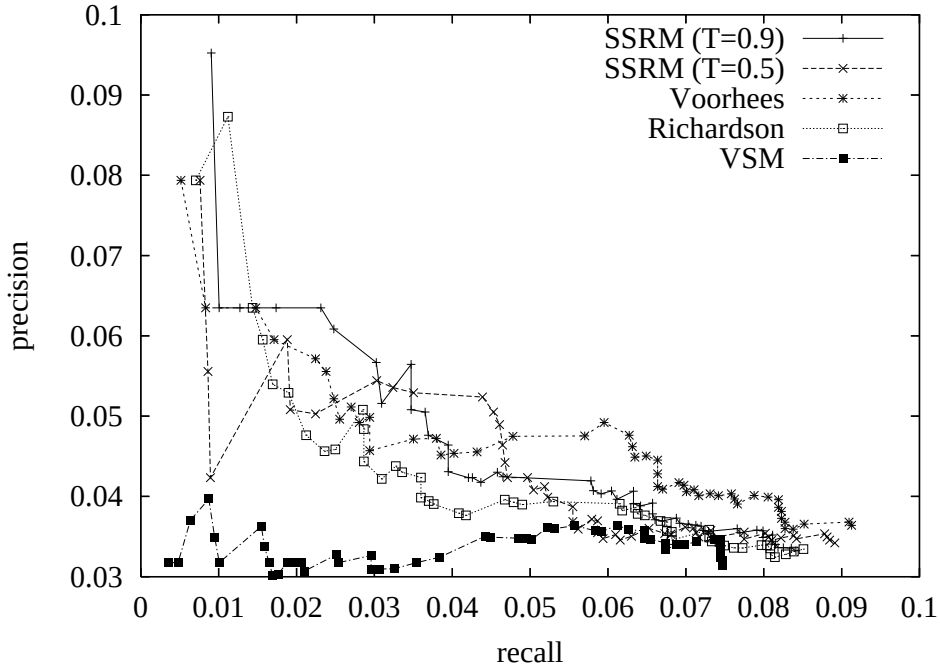

Fig. 6. Precision-recall diagram for retrievals on OHSUMED using MeSH.

ria. Several approaches to the problem of content-based image retrieval on the Web have been proposed and some have been implemented on research prototypes (e.g., ImageRover [36],WebSEEK [37], Diogenis [38]) and commercial systems (e.g., Google Image Search [16], Yahoo [17], Altavista [18]. Because, methods for extracting reliable and meaningful image content from Web pages by automated image analysis are not yet available images on the Web are typically described by text or attributes associated with images in `html` tags (e.g., filename, caption, alternate text etc.). These are automatically extracted from the Web pages and are used in retrievals. Google, Yahoo, and AltaVista are example systems of this category.

We choose the problem of image retrieval based on surrounding text as a case study for this evaluation. *SSRM* has been evaluated through *IntelliSearch* [19], a prototype Web retrieval system for Web pages and images in Web pages. An earlier system we built supported retrievals using only VSM [39]. In this work the system has been extended to support retrievals using *SSRM*

[16] `http://www.google.com/imghp`
[17] http://images.search.yahoo.com
[18] http://www.altavista.com/image
[19] http://www.intelligence.tuc.gr/intellisearch

with WordNet as the underlying reference ontology. The retrieval system of *IntelliSearch* is built upon Lucene[20] and the database stores more than 1.5 million web pages with images.

As it is typical in the literature [40, 39, 41], the problem of image retrieval on the Web is treated as one of text retrieval as follows: Images are described by the text surrounding them in the Web pages (i.e., captions, alternate text, image file names, page title). These descriptions are syntactically analyzed and reduced into term vectors which are matched against the queries. Similarly to the previous experiment, the similarity between a query and a document (image) is computed as

$$Sim(q,d) = Sim(q, d_{image\_file\_name}) + Sim(q, d_{caption}) + Sim(q, d_{page\_title}) + Sim(q, d_{alternate\_text}). \quad (8)$$

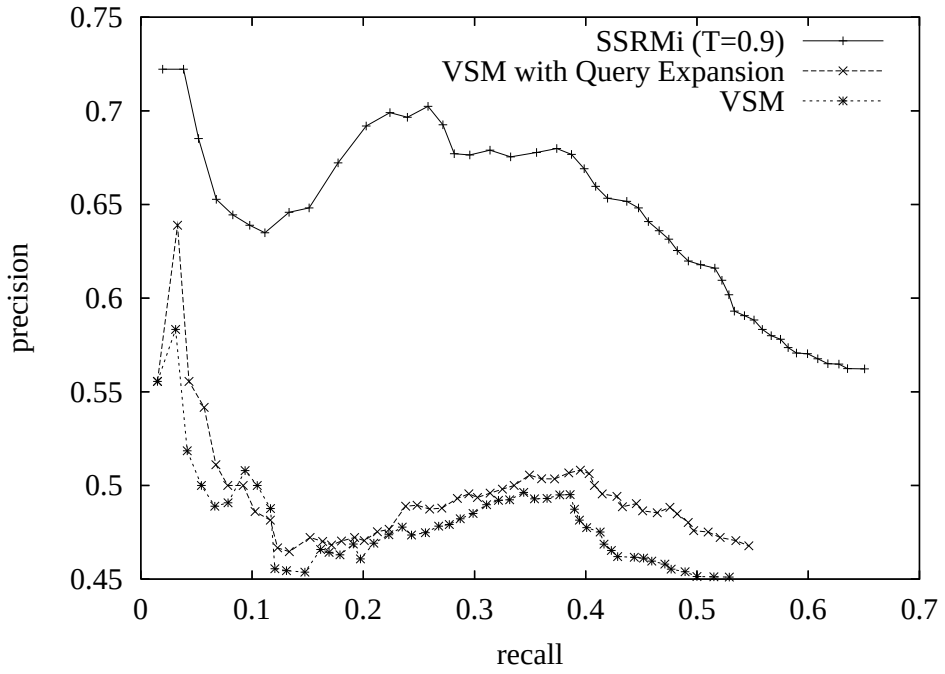

Fig. 7. Precision-recall diagram of *SSRM* and VSM for retrievals on the Web using WordNet.

For the evaluations, 20 queries were selected from the list of the most frequent Google image queries [21]. These are short queries containing between 1 and 4 terms. Each query retrieved

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

[21] http://images.google.com

the best 50 answers (the precision/recall plot of each method contains exactly 50 points). The evaluation is based on human relevance judgments by 5 human referees. Each referee evaluated a subset of 4 queries for both methods.

Fig. 7 indicates that *SSRM* is far more effective than VSM achieving up to 30% better precision and up to 20% better recall. A closer look into the results reveals that the efficiency of *SSRM* is mostly due to the contribution of non-identical but semantically similar terms. VSM (like most classical retrieval models relying on lexical term matching) ignore this information. In VSM, query terms may also be expanded with synonyms. Experiments with and without expansion by synonyms are presented. Notice that VSM with query expansion by synonyms improved the results of plain VSM only marginally, indicating that the performance gain of *SSRM* is not due to the expansion by synonyms but rather due to the contribution of semantically similar terms.

5 Conclusions

This paper makes two contributions. The first contribution is to experiment with several semantic similarity methods for computing the conceptual similarity between natural language terms using WordNet and MeSH. To our knowledge, similar experiments with MeSH have not been reported elsewhere. The experimental results indicate that it is possible for these methods to approximate algorithmically the human notion of similarity reaching correlation (with human judgment of similarity) up to 83% for WordNet and up to 74% for MeSH. The second contribution is SSRM, an information retrieval method that takes advantage of this result. *SSRM* outperforms VSM, the classic information retrieval method and demonstrates promising performance improvements over other semantic information retrieval methods in retrieval on OHSUMED, a standard TREC collection with medical documents which is available on the Web. Additional experiments have demonstrated the utility of SSRM in web image retrieval based on text image descriptions extracted automatically. *SSRM* has been also tested on Medline[22], the premier bibliographic database of the U.S. National Library of Medicine (NLM) [28]. All experiments

[22] http://www.nlm.nih.gov/databases/databases_medline.html

confirmed the promise of *SSRM* over classic retrieval models. *SSRM* can work in conjunction with any taxonomic ontology like MeSH or WordNet and any any associated document corpus. Current research is directed towards extending *SSRM* to work with compound terms (phrases), and more term relationships (in addition to the Is-A relationships).

Acknowledgement

Dr Qiufen Qi of Dalhousie University for prepared the MeSH terms and the queries for the experiments with MeSH and evaluated the results of retrievals on Medline. We thank Nikos Hurdakis, and Paraskevi Raftopoulou for valuable contributions into this work. The U.S. National Library of Medicine provided us with the complete data sets of MeSH and Medline. This work was funded by project MedSearch / BIOPATTERN (Fp6, Project No 508803) of the European Union (EU), the Natural Sciences and Engineering Research Council of Canada, and IT Interactive Services Inc.

References

1. R.B.-Yates, B.R.-Neto: Modern Information Retrieval. Addison Wesley Longman (1999)
2. Salton, G.: Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer. Addison-Wesley (1989)
3. Voorhees, E.: Query Expansion Using Lexical-Semantic Relations. In: ACM SIGIR'94, Dublin, Ireland (1994) 61–69
4. Richardson, R., Smeaton, A.: Using WordNet in a Knowledge-Based Approach to Information Retrieval. Techn. Report Working Paper: CA-0395, Dublin City University, Dublin, Ireland (1995)
5. Hersh, W., Buckley, C., Leone, T., Hickam, D.: OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. In: ACM SIGIR Conf. (1994) 192–201
6. Collins-Thomson, K., Callan, J.: Query Expansion Using Random Walk Models. In: ACM Conf. on Information and Knowledge Management (CIKM'05), Bremen, Germany (2005) 704–711
7. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. In: American Association for Artificial Intelligence (AAAI'06), Boston (2006)
8. Qiu, Y., Frei, H.: Concept Based Query Expansion. In: SIGIR Conf. on Research and Development in Information Retrieval, Pittsburgh, PA, MA (1993) 160–169
9. Mandala, R., Takenobu, T., Hozumi, T.: The Use of WordNet in Information Retrieval. In: COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, CA (1998) 469–477
10. Attar, R., Fraenkel, A.: Local Feedback in Full Text Retrieval Systems. Journal of the ACM **24**(3) (1977) 397–417

11. Liu, S., Liu, F., Yu, C., Meng, W.: An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. In: ACM SIGIR'04, Sheffied, Yorkshire, UK (2004) 266–272
12. Possas, B., Ziviani, N., Meira, W., Ribeiro-Neto, B.: Set-Based Vector Model: An Efficient Approach for Correlation-Based Ranking. ACM Trans. on Info. Systems **23**(4) (2005) 397–429
13. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and Application of a Metric on Semantic Nets. IEEE Trans. on Systems, Man, and Cybernetics **19**(1) (1989) 17–30
14. Wu, Z., Palmer, M.: Verb Semantics and Lexical Selection. In: Annual Meeting of the Associations for Computational Linguistics (ACL'94), Las Cruces, New Mexico (1994) 133–138
15. Li, Y., Bandar, Z.A., McLean, D.: An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. IEEE Trans. on Knowledge and Data Engineering **15**(4) (2003) 871–882
16. 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
17. Richardson, R., Smeaton, A., Murphy, J.: Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words. Techn. Report Working paper CA-1294, School of Computer Applications, Dublin City University, Dublin, Ireland (1994)
18. Lord, P., Stevens, R., Brass, A., Goble, C.: Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation. Bioinformatics **19**(10) (2003) 1275–83
19. Resnik, O.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. Journal of Artificial Intelligence Research **11** (1999) 95–130
20. Lin, D.: Principle-Based Parsing Without Overgeneration. In: Annual Meeting of the Association for Computational Linguistics (ACL'93), Columbus, Ohio (1993) 112–120
21. Jiang, J., Conrath, D.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: Intern. Conf. on Research in Computational Linguistics, Taiwan (1998)
22. Seco, N., Veale, T., Hayes, J.: An Intrinsic Information Content Metric for Semantic Similarity in WordNet. Techn. report, University College Dublin, Ireland (2004)
23. Tversky, A.: Features of Similarity. Psycological Review **84**(4) (1977) 327–352
24. Rodriguez, M., Egenhofer, M.: Determining Semantic Similarity Among Entity Classes from Different Ontologies. IEEE Trans. on Knowledge and Data Engineering **15**(2) (2003) 442–456
25. Varelas, G.: Semantic Similarity Methods in Wordnet and their Application Information Retrieval on the Web. Dissertation thesis, Dept. of Electronic and Comp. Engineering, Technical Univ. of Crete (TUC), Chania, Greece (2005) http://www.intelligence.tuc.gr/publications/Varelas.pdf.
26. Miller, G., Charles, W.: Contextual Correlates of Semantic Similarity. Language and Cognitive Processes **6** (1991) 1–28
27. 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
28. Hliaoutakis, A., Varelas, G., Petrakis, E., Milios, E.: MedSearch: A Retrieval System for Medical Information Based on Semantic Similarity. In: European Conf. on Research and Advanced Technology for Digital Libraries (ECDL'06), Alicante, Spain (2006)

29. Rocchio, J.: Relevance Feedback in Information Retrieval. In Salton, G., ed.: The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice Hall, Englewood Cliffs (1971) 313–323
30. Rui, Y., Huang, T.S., Ortega, M., Mechrota, S.: Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. IEEE Trans. on Circ. and Syst. for Video Techn. **8**(5) (1998) 644–655
31. Patwardhan, S., Banerjee, S., Petersen, T.: Using Measures of Semantic Relatedness for Word Sense Disambiguation. In: Intern. Conf. on Intelligent Text Processing and Comutational Linguistics, Mexico City (2003) 17–21
32. Arasu, A., Cho, J., Garcia-Molina, H., Paepke, A., Raghavan, S.: Searching the Web. ACM Transactions on Internet Technology **1**(1) (2001) 2–43
33. R. Baeza-Yates, E.: Modern Information Retrieval. Addison Wesley (1999)
34. Kherfi, M., Ziou, D., Bernardi, A.: Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. ACM Computing Surveys **36**(1) (2004) 35–67
35. Smeulders, A.W., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(11) (2000) 1349–1380
36. Taycher, L., Cascia, M., Sclaroff, S.: Image Digestion and Relevance Feedback in the ImageRover WWW Search Engine. In: 2^{nd} Intern. Conf. on Visual Information Systems, San Diego (1997) 85–94
37. Smith, J., Chang, S.F.: Visually Searching the Web for Content. IEEE Multimedia **4**(3) (1997) 12–20
38. Aslandongan, Y., Yu, C.: Evaluating Strategies and Systems for Content-Based Indexing of Person Images on the Web. In: 8^{th} Intern. Conf. on Multimedia, Marina del Rey, CA (2000) 313–321
39. Voutsakis, E., Petrakis, E., Milios, E.: Weighted Link Analysis for Logo and Trademark Image Retrieval on the Web. In: IEEE/WIC/ACM Intern. Conf. on Web Intelligence (WI2005), Compiegne University of Technology, France (2005)
40. Shen, H.T., Ooi, B.C., Tan, K.L.: Giving Meanings to WWW Images. In: 8^{th} Intern. Conf. on Multimedia, Marina del Rey, CA (2000) 39–47
41. Voutsakis, E., Petrakis, E., Milios, E.: IntelliSearch: Intelligent Search for Images and Text on the Web. In: Intern. Conf. on Image Analysis and Recognition (ICIAR'06), Povoa de Varzim (2006)

Table of Contents

List of Figures

List of Tables