Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web

Giannis Varelas† varelas@intelligence.tuc.gr
Epimenidis Voutsakis† pimenas@softnet.tuc.gr
Paraskevi Raftopoulou† paraskevi@intelligence.tuc.gr
Euripides G.M. Petrakis† petrakis@intelligence.tuc.gr
Evangelos E. Milios‡ eem@cs.dal.ca

† Dept. of Electronic and Comp. Engineering, Technical University of Crete (TUC), Chania, Crete, Greece
‡ Faculty of Comp. Science, Dalhousie University, Halifax, Nova Scotia, Canada

ABSTRACT

Semantic Similarity relates to computing the similarity between concepts which are not lexicographically similar. We investigate approaches to computing semantic similarity by mapping terms (concepts) to an ontology and by examining their relationships in that ontology. Some of the most popular semantic similarity methods are implemented and evaluated using WordNet as the underlying reference ontology. Building upon the idea of semantic similarity, a novel information retrieval method is also proposed. This method is capable of detecting similarities between documents containing semantically similar but not necessarily lexicographically similar terms. The proposed method has been evaluated in retrieval of images and documents on the Web. The experimental results demonstrated very promising performance improvements over state-of-the-art information retrieval methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models, Query Formulation,Search Process*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Dictionaries, Thesauruses*

General Terms

Performance, Experimentation, Algorithms

Keywords

Information Retrieval, Semantic Similarity, WordNet, World Wide Web

WIDM'05, November 5, 2005, Bremen, Germany.

1. INTRODUCTION

Information retrieval is currently being applied in a variety of application domains from database systems to Web information search engines. The main idea is to locate documents that contain terms that the users specify in queries. The lack of common terms in two documents does not necessarily mean that the documents are not related. Retrieval, by classical information retrieval models (e.g., Vector Space, Probabilistic, Boolean) [14] is based on lexicographic term matching. However, two terms can be semantically similar (e.g., can be synonyms or have similar meaning) although they are lexicographically different. Therefore, retrieval by classical retrieval methods will fail to retrieve documents with semantically similar terms. This is exactly the problem this work is addressing.

In the first part of this work we propose discovering semantically similar terms using WordNet [1]. Several methods have been implemented and evaluated. In the second part of this work we propose the *Semantic Similarity Retrieval Model* (SSRM), a general document similarity and information retrieval method suitable for retrieval in conventional document collections and the Web. 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 by applying a semantic query in the neighborhood of each term) 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.

The term-based Vector Space Model (VSM) [20] (the state-of-the-art document retrieval method) and SSRM (our proposed retrieval method), have been implemented and evaluated on a retrieval system for images and documents on the Web [26]. The system stores a crawl of the Web with more that 1,5 million Web pages with images. SSRM demonstrated very promising performance achieving significantly better precision and recall than VSM.

The contributions of the proposed work are summarized in the following:

- A framework and a system for evaluating the performance of several semantic similarity methods in Word-

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

Net is implemented. The system is available on the Web [2].

- SSRM, a novel information retrieval model based on the integration of semantic similarity methods in document matching is proposed.
- SSRM and VSM have been evaluated and integrated into a fully automated information retrieval method for Web pages and images in Web pages. This system is also available on the Web [3].

The rest of this paper in organized as follows: Semantic similarity methods and their application to the WordNet lexical ontology are discussed in Sec. 2. The proposed semantic similarity retrieval model is presented in Sec. 3. A prototype Web retrieval system integrating SSRM is presented in Sec. 4. Experimental results are presented in Sec. 5 followed by conclusions in Sec. 6.

2. WORDNET AND SEMANTIC SIMILARITY METHODS

WordNet [4] 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 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.

It is commonly argued that language semantics are mostly captured by nouns (and noun phrases) so that it is common to built retrieval methods based on noun representations extracted from documents and queries. In the following, we only use the nouns and the Hyponym/Hypernym relationships from WordNet.

Several methods for determining semantic similarity between terms have been proposed in the literature and most of them have been tested on WordNet [5]. Similarity measures apply only for nouns and verbs in WordNet (taxonomic properties for adverbs and adjectives do not exist). 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, 27, 5, 4].

Information Content Methods: Measure the difference in information content of the two terms as a function of their probability of occurrence in a corpus [8, 15, 6, 3]. In this work information content is computed according to [21]: WordNet is used as a statistical resource for computing the probabilities of occurrence of terms. 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.

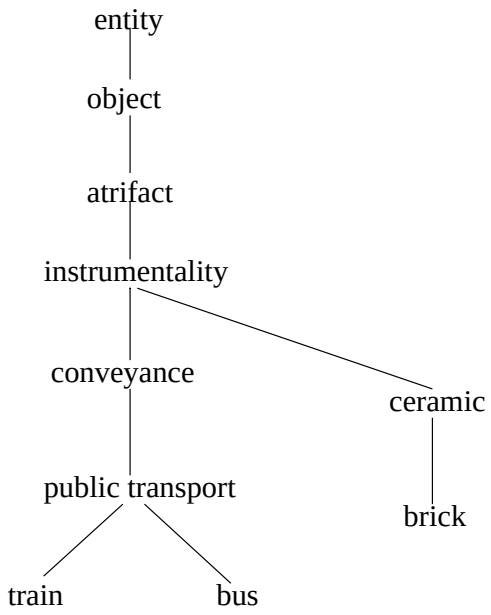

Figure 1: A fragment of the WordNet Is-A hierarchy.

Feature based Methods: Measure the similarity between two terms as a function of their properties (e.g., their definitions or "glosses" in WordNet) or based on their relationships to other similar terms in the taxonomy [23].

Hybrid methods combine the above ideas [19].

Semantic similarity methods can also be distinguished between:

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

Cross Ontology similarity methods for comparing terms from two different ontologies (e.g., WordNet and MeSH [6], an ontology for medical terms developed by the U.S. National Library of Medicine).

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).

[2]http://www.ece.tuc.gr/similarity
[3]http://www.ece.tuc.gr/intellisearch
[4]http://wordnet.princeton.edu
[5]http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi
[6]http://www.nlm.nih.gov/mesh/meshhome.html

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 methods.

The focus of this work is on single ontology methods. All methods above are implemented and integrated into a semantic similarity system which is available on the Web [7]. Fig. 2 illustrates the architecture of this system. The system communicates with WordNet 2.0. Each term is represented by its tree hierarchy (corresponding to an XML file) which is stored in the XML repository. These XML files are created using the WordNet XML Web-Service[8]. 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).

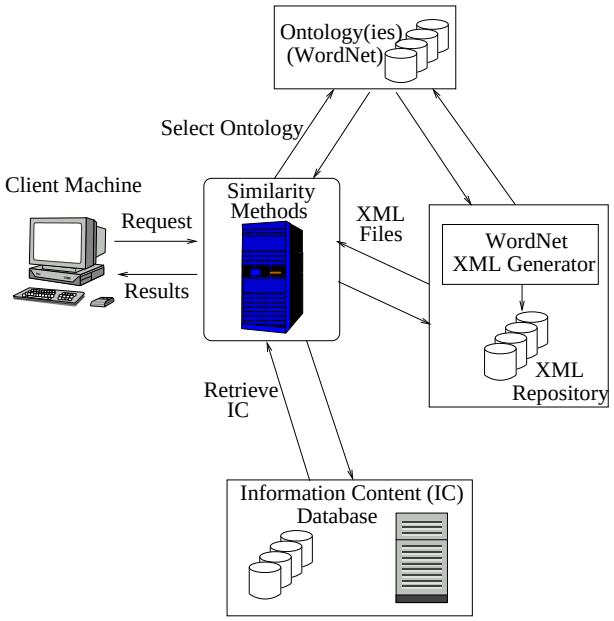

Figure 2: Semantic Similarity System.

3. SEMANTIC SIMILARITY RETRIEVAL MODEL (SSRM)

Queries and documents are syntactically analyzed and reduced into term (noun) vectors. A term is usually defined as a stemmed non stop-word. 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 [20] is used for computing the weight. Typically, the weight d_i of a term i in a document is computed as

$$d_i = tf_i \cdot idf_i, \tag{1}$$

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

Traditionally, the similarity between two documents (e.g., a query q and a document d) is computed according to the Vector Space Model (VSM) [20] as the cosine of the inner product between their document vectors

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

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 *bag of words* model and is the state of the art model for document retrieval.

The lack of common terms in two documents does not necessarily mean that the documents are unrelated. Similarly, relevant documents may not contain the same terms. 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").

We propose discovering semantically similar terms using WordNet and semantic similarity methods. The evaluation of the semantic similarity methods indicate that the method by [5] is particularly effective, achieving up to 82% correlation with results obtained by humans (Sec. 5.1).

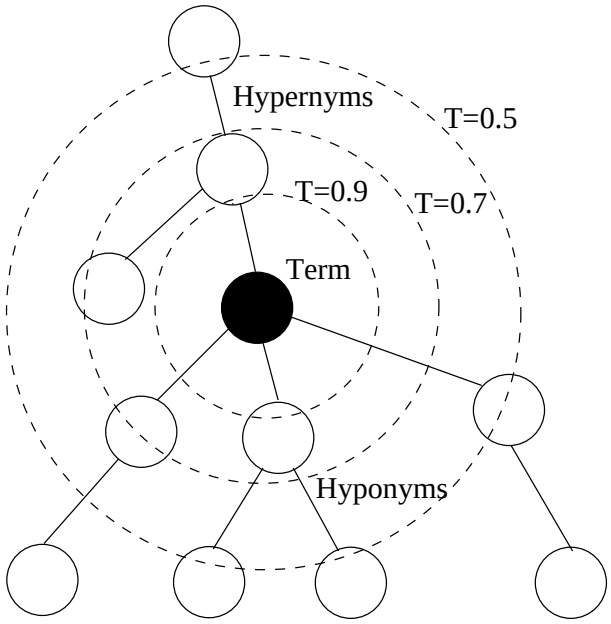

Figure 3: Term expansion using WordNet.

In the sequel, SSRM works in three steps:

Term 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

$$q_i = q_i + \sum_{\text{sim}(i,j)\geq t}^{j\neq i} q_j \text{sim}(i,j), \tag{3}$$

[7] http://www.ece.tuc.gr/similarity
[8] http://wnws.sourceforge.net

where t is a user defined threshold ($t = 0.8$ in this work). This formulae suggests assigning higher weights to semantically similar terms within the query (e.g., "railway", "train", "metro"). The weights of non-similar terms remain unchanged (e.g., "train", "house").

Term Expansion: First, the query is augmented by synonym terms (the most common sense is taken). Then, the query is augmented by hyponyms and hypernyms which are semantically similar to terms already in the query. Fig. 3 illustrates this process: Each query term is represented by its WordNet tree hierarchy. The neighborhood of the term 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

$$q_i = \begin{cases} \sum_{\text{sim}(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j \text{sim}(i,j), & i \text{ is a new term} \\ q_i + \sum_{\text{sim}(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j \text{sim}(i,j), & i \text{ had weight } q_i, \end{cases} \quad (4)$$

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. 4 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.

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

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

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].

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 finally 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). In this work work $T = 0.9$ (the query is expanded only with very similar terms). The specification of T requires further investigation (e.g., appropriate threshold values can be learned by training). Word sense disambiguation [11] can also be applied to detecting the correct sense to expand (rather than expanding the most common sense of each term).

3.1 Discussion

WordNet has been used many times in information retrieval research with unsatisfactory results in most cases. Voorhees [24] proposed expanding query terms with synonyms, hyponyms and hypernyms 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 [16] proposed taking the summation of the semantic similarities between all possible combinations of document and query terms. However, they ignored the relevant significance of terms (as captured by $tf \cdot idf$ weights) and neither they considered term expansion nor re-weighting. Their method degraded the performance of retrievals. Our proposed method combines ideas from both these methods, takes term weights into account, introduces a analytic and intuitive term expansion and re-weighting method (as opposed to the ad-hoc method by Voorhees [24]) and suggests a document similarity formulae that takes the above information into account.

Similarly to VSM, our proposed model allows for non-binary weights in queries and in documents (initial weights are computed using the standard $tf \cdot idf$ formulae). The model also allows 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).

SSRM relaxes the requirement of classical retrieval models that conceptually similar terms are mutually independent (known also as "synonymy problem"). It takes into account all possible 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. 2 of VSM. SSRM approximates VSM in the case of non-semantically similar terms: If $\text{sim}(i,j) = 0\ \forall i \neq j$ then Eq. 5 approximates Eq. 2 (the two formulae become identical except the normalization factors). In this case, the similarity between two documents is computed as a function of weight similarities between identical terms.

Expanding and re-weighting is fast for queries (queries are short in most cases specifying 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 in the document and in 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 [18]. The proposed method of query expansion and term re-weighting resembles also approaches which attempt to improve the query with terms which are obtained from a similarity thesaurus (e.g., based on term to term relationships [12, 9]) which is usually extracted by automatic or semi-automatic corpus analysis (global analysis). A thesaurus would not only add new terms to SSRM but also reveal new relationships not existing in Wordnet. This approach is expensive in time and also depends on the corpus.

Our proposed approach is independent of the corpus and works by discovering term associations based on their conceptual similarity in WordNet (or in a lexical ontology specific to the application domain at hand), it is faster and

more intuitive. The proposed query expansion scheme is also complementary to methods which expand the query with co-occurrent terms (e.g., "railway", "station") in retrieved documents [1] (local analysis). Expansion with co-occurrent terms (the same as a thesaurus like expansion) can be introduced as additional expansion step in the method. Finally, SSRM needs to be extended to work with phrases in addition to single word terms [7].

4. WEB RETRIEVAL SYSTEM

The proposed method has been evaluated using a prototype retrieval system for images in Web pages [26]. The system is available on the Web [9]. Fig. 4 illustrates the architecture of the system. The system consists of several modules, the most important of them being the following:

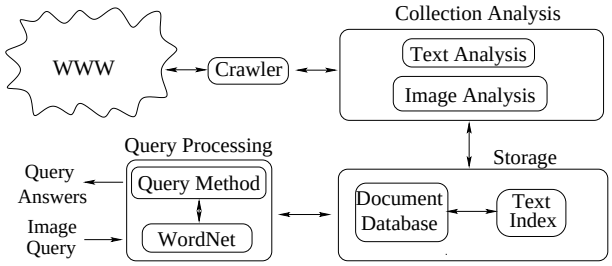

Figure 4: Web System Architecture.

Crawler module: Implemented based upon Larbin [10], the crawler assembled locally a collection of 1,5 million pages with images. The crawler started its recursive visit of the Web from a set of 14,000 pages which is assembled from the answers of Google image search [11] to 20 queries on various topics (e.g., topics related to Linux and software products). The crawler worked recursively in breadth-first order and visited pages up to a depth of 5 links from each origin.

Collection Analysis module: The content of crawled pages is analyzed. Text, images, link information (forward links) and information for pages that belong to the same site is extracted. For each image, its text description is extracted.

Storage module: Implements storage structures and indices providing fast access to Web pages and information extracted from them (i.e., text, image descriptions and link information).

Query Processing module: Queries are issued by keywords or free text.

The database is implemented in BerkeleyDB [12]. Two inverted files implement the connectivity server [2] and provide fast access to linkage information between pages (backward and forward links). Two inverted files associate terms with their intra and inter document frequencies and allow for fast computation of term vectors.

[9]http://www.ece.tuc.gr/intellisearch
[10]http://larbin.sourceforge.net
[11]http://www.google.com/imghp
[12]http://www.sleepycat.com

The system is designed to support queries by image content for logo and trademark images on the Web. As it is typical in the literature [22], the problem of image retrieval on the Web is treated as one of text retrieval as follows: Images are described by text surrounding the images in the Web pages (i.e., captions, alternate text, image file names, page title). These descriptions are syntactically analyzed and reduced into vectors of stemmed terms (nouns) which are matched against the queries.

In [26] the system supported retrievals using only VSM. In this work the system is extended to support retrievals using our proposed Semantic Similarity Retrieval Model in addition to VSM. The user is prompted at the interface to select the desired retrieval method.

5. EXPERIMENTAL RESULTS

In the first part of this section we present results on the evaluation of the similarity methods described in Sec. 2. The second part presents results obtained from the application of SSRM to the problem of information retrieval on the Web.

5.1 Evaluation of Semantic Similarity Methods

In the following we present a comparative evaluation of various semantic similarity methods. In accordance with previous research, we evaluated the results obtained by applying the semantic similarity methods of Sec. 2 and by correlating their similarity scores with the scores obtained by human judgments in the experiment by Miller and Charles [10]: 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. 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 this experiment, we implemented several similarity measures reported in the literature and compared the computed similarity scores for the same terms as in Miller and Charles with the human relevance results reported there. The higher the correlation of a method the better the method is (i.e., the more it approaches the results of human judgements).

Table 1 shows the correlation obtained by each method. Jiang and Conrath [3] 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. [5] is among the best and it is also the fastest. These results lead to the following observations:

- All Information Content methods perform very well and close to the upper bound suggested by Resnik [15].
- Methods that consider the positions of the terms in the hierarchy (e.g., [5]), 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 (they do not fully exploit this information). However, Hybrid and feature based methods (e.g., [19]) are mainly targeted towards cross ontology similarity

Table 1: Evaluation of Edge Counting, Information Content, Feature based and Hybrid semantic similarity methods.

Method	Type	Correlation
Rada [13]	Edge Counting	0.59
Wu [27]	Edge Counting	0.74
Li [5]	Edge Counting	0.82
Leacock [4]	Edge Counting	0.82
Richardson [17]	Edge Counting	0.63
Resnik [15]	Information Content	0.79
Lin [6]	Information Content	0.82
Lord [8]	Information Content	0.79
Jiang [3]	Information Content	0.83
Tversky [23]	Feature	0.73
Rodriguez [19]	Hybrid	0.71

applications where edge counting and information content methods do not apply.

5.2 Evaluation of Retrieval Methods

We choose the problem of image retrieval based on surrounding text as a case study for the evaluation of the proposed approach. The following methods are evaluated:

Semantic Similarity Retrieval Model (SSRM): The proposed method. To avoid topic drift only very similar terms are included in the expansion step: Each query term is expanded with synonyms and semantic similar terms with similarity greater than $T = 0.9$. Semantic similarity between terms is computed by [5].

Vector Space Model (VSM) [20]: The state-of-the-art text retrieval method. Text queries are also augmented by synonyms.

For the evaluations, 20 queries were selected from the list of the most frequent Google image queries [13]. These are rather short queries containing between 1 and 4 terms. The evaluation is based on human relevance judgments by 5 human referees. Each referee evaluated a subset of 4 queries for both methods.

To evaluate the effectiveness of each candidate method, the following quantities are computed:

Precision that is, the percentage of relevant images retrieved with respect to the number of retrieved images.

Recall that is, the percentage of relevant images retrieved with respect to the total number of relevant images in the database. Due to the large size of the data set, it is practically impossible to compare every query with each database image. To compute recall, for each query, the answers obtained by all candidate methods are merged and this set is considered to contain the total number of correct answers. This is a valid sampling method known as *"pooling method"* [25]. This method allows for relative judgements (e.g., method A retrieves 10% more relevant answers than method B) but does not allow for absolute judgements (e.g., method A retrieved 10% of the total relevant answers).

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

Each method is represented by a *precision-recall* curve. Each query retrieves the best 50 answers and each point on a curve is the average precision and recall over 20 queries. Precision and recall values are computed from each answer set after each answer (from 1 to 50) 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 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 recall.

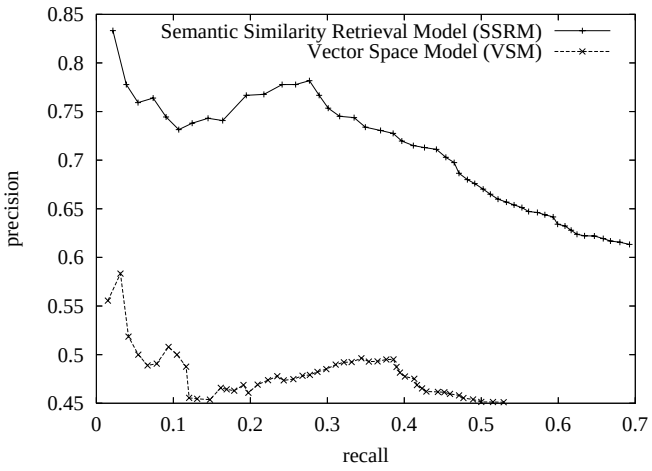

Figure 5: Precision-recall diagram of SSRM and VSM.

Fig. 5.2 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 lexicographic term matching) ignore this information.

6. CONCLUSIONS

We experimented with several semantic similarity methods for computing the conceptual similarity between natural language terms using WordNet. The experimental results indicate that it is possible for these methods to approximate algorithmically the human notion of similarity reaching correlation up to 83%. Based on this observation, we demonstrated that it is possible to exploit this information (as embedded in taxonomic ontologies and captured algorithmically by semantic similarity methods) for improving the performance of retrievals on the Web. For this purpose, the Semantic Similarity Retrieval Model (SSRM), a novel document retrieval model that incorporates conceptual similarity into its retrieval mechanism is proposed and evaluated as part of this work. SSRM can work in conjunction with any taxonomic ontology (e.g., application specific ontologies). The evaluation demonstrated very promising performance improvements over the Vector Space Model (VSM), the state-of-the-art document retrieval method.

Future work includes experimentation with more data sets (e.g., TREC, Medline) and ontologies (e.g., the MeSH ontology of medical terms) and experimentation with more

application domains (e.g., document clustering, document searching in P2P systems). SSRM can also be extended to work with compound terms (phrases) and terms with different part of speech (in addition to single word nouns), more term relationships in WordNet (in addition to the Is-A relationships) and with terms and term relationships not existing in WordNet (e.g., obtained from a thesaurus). Also, more elaborate query expansion methods (e.g., methods for specifying thresholds for query expansion) need to be investigated.

7. REFERENCES

[1] R. Attar and A. Fraenkel. Local Feedback in Full Text Retrieval Systems. *Journal of the ACM*, 24(3):397–417, 1977.

[2] K. Bharat, A. Broder, M. R. Henzinger, P. Kumar, and S. Venkatasubramanian. The Connectivity Server: Fast Access to Linkage Information on the Web. In *Intern. World Wide Web Conference (WWW-7)*, pages 469–477, Brisbane, Australia, 1998.

[3] J. Jiang and D. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Intern. Conf. on Research in Computational Linguistics*, Taiwan, 1998.

[4] C. Leacock and M. Chodorow. Combining Local Context and WordNet Similarity for Word Sense Identification in WordNet. In C. Fellbaum, editor, *An Electronic Lexical Database*, pages 265–283. MIT Press, 1998.

[5] Y. Li, Z. A. Bandar, and D. McLean. An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Trans. on Knowledge and Data Engineering*, 15(4):871–882, July/Aug. 2003.

[6] D. Lin. Principle-Based Parsing Without Overgeneration. In *Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 112–120, Columbus, Ohio, 1993.

[7] S. Liu, F. Liu, C. Yu, and W. Meng. An Effective Approach to Document Retrieval via Utilizing WordNet and Recognizing Phrases. In *ACM SIGIR'04*, pages 266–272, Sheffied, Yorkshire, UK, 2004.

[8] P. Lord, R. Stevens, A. Brass, and C. Goble. Investigating Semantic Similarity Measures across the Gene Ontology: the Relationship between Sequence and Annotation. *Bioinformatics*, 19(10):1275–83, 2003.

[9] R. Mandal, T. Takenobu, and T. Hozumi. The Use of WordNet in Information Retrieval. In *COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, pages 469–477, Montreal, CA, 1998.

[10] G. Miller and W. Charles. Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6:1–28, 1991.

[11] S. Patwardhan, S. Banerjee, and T. Petersen. Using measures of semantic relatedness for word sense disambiguation. In *Intern. Conf. on Intelligent Text Processing and Comutational Linguistics*, pages 17–21, Mexico City, 2003.

[12] Y. Qiu and H. Frei. Concept Based Query Expansion. In *SIGIR Conf. on Research and Development in Information Retrieval*, pages 160–169, Pittsburgh, PA, MA, 1993.

[13] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and Application of a Metric on Semantic Nets. *IEEE Trans. on Systems, Man, and Cybernetics*, 19(1):17–30, Jan./Feb. 1989.

[14] R.B.-Yates and B.R.-Neto. *Modern Information Retrieval*. Addison Wesley Longman, 1999.

[15] O. Resnik. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity and Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.

[16] R. Richardson and A. Smeaton. Using WordNet in a Knowledge-Based Approach to Information Retrieval. Techn. Report Working Paper: CA-0395, Dublin City University, Dublin, Ireland, 1995.

[17] R. Richardson, A. Smeaton, and J. Murphy. 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] J. Rocchio. Relevance Feedback in Information Retrieval. In G. Salton, editor, *The SMART Retrieval System - Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, Englewood Cliffs, 1971.

[19] M. Rodriguez and M. Egenhofer. Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Trans. on Knowledge and Data Engineering*, 15(2):442–456, March/April 2003.

[20] G. Salton. *Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer*. Addison-Wesley, 1989.

[21] N. Seco, T. Veale, and J. Hayes. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. Techn. report, University College Dublin, Ireland, 2004.

[22] H.-T. Shen, B.-C. Ooi, and K.-L. Tan. Giving Meanings to WWW Images. In 8^{th} *Intern. Conf. on Multimedia*, pages 39–47, Marina del Rey, CA, 2000.

[23] A. Tversky. Features of Similarity. *Psycological Review*, 84(4):327–352, 1977.

[24] E. Voorhees. Query Expansion Using Lexical-Semantic Relations. In *ACM SIGIR'94*, pages 61–69, Dublin, Ireland, 1994.

[25] E. Voorhees and D. Harmann. Overview of the Seventh Text REtrieval Conference (TREC-7). In *NIST Special Publication 500-242: The Seventh Text REtrieval Conference (TREC-7)*, pages 1–23, 1998. http://trec.nist.gov/pubs/trec7/t7_proceedings.html.

[26] E. Voutsakis, E. Petrakis, and E. Milios. 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.

[27] Z. Wu and M. Palmer. Verb Semantics and Lexical Selection. In *Annual Meeting of the Associations for Computational Linguistics (ACL'94)*, pages 133–138, Las Cruces, New Mexico, 1994.