International Journal of Computer Applications Technology and Research
Volume 2– Issue 1, 77-80, 2013, ISSN: 2319–8656

Semantic Information Retrieval based on Wikipedia Taxonomy

May Sabai Han
University of Technology
Yatanarpon Cyber City
Myanmar

Abstract: Information retrieval is used to find a subset of relevant documents against a set of documents. Determining semantic similarity between two terms is a crucial problem in Web Mining for such applications as information retrieval systems and recommender systems. Semantic similarity refers to the sameness of two terms based on sameness of their meaning or their semantic contents. Recently many techniques have introduced measuring semantic similarity using Wikipedia, a free online encyclopedia. In this paper, a new technique of measuring semantic similarity is proposed. The proposed method uses Wikipedia as an ontology and spreading activation strategy to compute semantic similarity. The utility of the proposed system is evaluated by using the taxonomy of Wikipedia categories.

Keywords: information retrieval; semantic similarity; spreading activation strategy; wikipedia taxonomy; wikipedia categories

1. INTRODUCTION

Information in WWW are scattered and diverse in nature. So, users frequently fail to describe the information desired. Traditional search techniques are constrained by keyword based matching techniques. Hence low precision and recall is obtained [2]. Many natural language processing applications must estimate the semantic similarity of pairs of text fragments provided as input, e.g. information retrieval, summarization, or textual entailment. A simple lexical overlap measure cannot be successful when text similarity is not based on identical words and in general when words are not independent [3].

It has long been recognized that in order to process natural language, computers require access to vast amount of common-sense and domain-specific world knowledge. However, prior work on semantic relatedness was based on purely statistical techniques that did not make use of background knowledge or on lexical resources that incorporate very limited knowledge about the world [1].

Many natural language processing tasks require external sources of lexical semantic knowledge such as Wordnet. Traditionally, these resources have been built manually by experts in a time consuming and expensive manner [4].

An advantage of using the "ontology" approach, whether based on a designed or emergent ontology, is that the terms can be explicitly linked or mapped to semantic concepts in other ontologies, and are thus available for reasoning in more sophisticated language understanding systems. Using the traditional approach of a controlled, designed ontology has many disadvantages beginning with the often difficult task of designing and implementing the ontology. Once that it done, it must be maintained and modified, an important process in domains where the underlying concepts are evolving rapidly [5].

Wikipedia has recently provided a wide range of knowledge including some special proper nouns in different areas of expertise (e.g., Obama) which is not described in WordNet. It also includes a large volume of articles about almost every entity in the world. Wikipedia provides a semantic network for computing semantic relatedness in a more structured fashion than a search engine and with more coverage than WordNet. And Wikipedia articles have been categorized by providing a taxonomy, categories. This feature provides the hierarchical structure or network. Wikipedia also provides articles link graph. So many researches has recently used Wikipedia as an ontology to measure semantic similarity.

We propose a method to use structured knowledge extracted from the English version of Wikipedia to compute semantic similarity. This model takes the system of categories in Wikipedia as a semantic network by considering that every article in Wikipedia as a concept. Our system uses spreading activation strategy on the network of Wikipedia categories to evaluate semantic similarity.

The rest of the paper is organized as follows. Section 2 expresses about information retrieval based on semantic similarity. Section 3 describes motivation for the proposed system. Section 4 discusses related semantic similarity computing techniques based on Wikipedia. Section 5 provides framework of our proposed system. Section 6 mentions about semantic similarity computing using spreading activation strategy, and section 7 concludes.

1.1 Spreading Activation Strategy

Spreading Activation Strategy is a technique that has been widely adopted for associative retrieval. In associative retrieval, the idea is that it is possible to retrieve relevant documents if they are associated with other documents that have been considered relevant by the user. Also it has proved a significant result in word sense disambiguation. In Wikipedia the links between categories show association between concepts of articles and hence can be used as such for finding related concepts to a given concept. The algorithm starts with a set of activated nodes and, in each iteration, the activation of nodes is spread to associated nodes. The spread of activation may be directed by addition of different constraints like distance constraints, fan out constraints, path constraint, threshold. These parameters are mostly domain specific [5].

2. INFORMATION RETRIEVAL BASED ON SEMANTIC SIMILARITY

Information retrieval (IR) is the task of representing, storing, organizing, and offering access to information items. IR is different from data retrieval, which is about finding precise data in databases with a given structure. In IR systems, the information is not structured; it is contained in free form in text (webpages or other documents) or in multimedia content. The first IR systems implemented in 1970's were designed to work with small collections of text (for example legal documents). Some of these techniques are now used in search engines. The aim is to retrieve all the relevant information according to the given query.

There is a huge quantity of text, audio, video, and other documents relating to the various subjects available on the Internet. With the explosive growth of information, it is becoming increasingly difficult to retrieve the relevant documents. This begins challenges to IR community and motivate researcher to look for information retrieval system which can retrieve information based on some higher level of understanding of query. This higher level of understanding can only be achieved through processing of text based on semantics, which is not possible by considering a document as a "bag of words". So, nowadays, several semantic similarity techniques have been used in information retrieval systems.

The semantic similarity computing techniques define how to compare query requests to the collection of documents to obtain the semantically related documents based on the concept of using ontology. Semantic similarity computing methods have to calculate the relatedness of two concepts though they don't have the exact match. Therefore, the percentage of relevant information we get mainly depends on the semantic similarity matching function we used. For the above fact, more and more semantic similarity methods are discovered to produce the most semantically related results.

3. MOTIVATION

Vector space model represents a document or a query as a vector. Although the term vector similarity computing is applied in a number of such applications for its simplicity and reasonable accuracy, it has a problem of lack of semantic. This is due to the representation of document in a linear form (i.e., a vector of features) in which semantic relations among features are ignored. An example for such problem is found in recommender systems which find people with similar preference according to their old transactions. Therefore several approaches have developed to enhance semantic similarity distance. Some approaches use the ontology to construct the taxonomy of concepts and relations for the fragments to be compared. Building and maintaining those knowledge bases require a lot of effort from expert. Moreover, only the domain specific terms or a small fraction of the vocabulary of a language are covered by the bases. Wikipedia provides a knowledge base for computing word relatedness in a more structured fashion than a search engine and with more coverage than WordNet. So, the idea of using Wikipedia is intended for computing semantic similarity in the proposed system.

4. RELATED WORK

The depth and coverage of Wikipedia has received a lot of attention from researchers who have used it as a knowledge source for computing semantic relatedness.

Explicit Semantic Analysis (ESA) [1] represents the meaning of texts in a high-dimensional space of concepts derived from Wikipedia. ESA uses machine learning techniques to explicitly represent the meaning of any text as a weighted vector of Wikipedia-based concepts. Assessing the relatedness of texts in this space amounts to comparing the corresponding vectors using conventional metrics (e.g., cosine). However, ESA does not use link structure and other structures knowledge from Wikipedia, although these contain valuable information about relatedness between articles.

Milne and Witten [9] measure semantic relatedness by using hyperlink structure of Wikipedia. Each article is represented by a list of its incoming and outgoing links. To compute relatedness, they use tf-idf using link counts weighted by the probability of each link occurring.

In WikiRelate [11], the two articles corresponding to two terms are retrieved firstly. Then the categories related to these articles are extracted and map onto the category network. Given the set of paths found between the category pairs, Strube and Ponzetto compute the relatedness by selecting the shortest path and the path which maximizes information content for information content based measures.

WikiWalk[10] evaluates methods for building the graph, including link selection strategies and performing random walks based on Personalized PageRank to obtain stationary distributions that characterize each text. Senamtic relatedness is computed by comparing the distributions.

Majid Yazdani et al. [3] build a network of concepts from Wikipedia documents using a random walk approach to compute distances between documents. Three algorithms for distance computation such as hitting/commute time, personalized page rank, and truncated visiting probability are proposed. Four types of weighted links in the document network such as actual hyperlinks, lexical similarity, common category membership and common template use are considered. The resulting network is used to solve three benchmark semantic tasks- word similarity, paraphrase detection between sentences, and document similarity by mapping pairs of data to the network, and then computing a distance between these representations.

Behanam et al. [8] extracted the multi-tree for each entity from Wikipedia categories network. Then combined two multi-trees and used multi-tree similarity algorithm to this combined tree to compute similarity.

Lu Zhiqiang et al. [6] used snippets from Wikipedia to calculate the semantic similarity between words by using cosine similarity and TF-IDF. That is different from other methods which used Wikipedia taxonomy. The stemmer algorithm and stop words are also applied in the preprocessing the snippets from Wikipedia.

In [5], Wikipedia articles, and the category and article link graphs are used to predict concepts common to a set of documents. Zareen Saba Syed et al. describe several algorithms to aggregate and refine results, including the use of spreading activation to select the most appropriate terms.

Stephan Gouws et al. [12] propose the Target Activation Approach(TAA) and the Agglomerative Approach (AA) for computing semantic relatedness by spreading activation energy over the hyperlink structure of Wikipedia. Relatedness between two nodes can be measured as either 1) the ratio of initial energy that reaches the target node, or 2) the amount of overlap between their individual activation vectors by spreading from both nodes individually. The second method is

adaptation of the Wikipedia Link-based Measure (WLM) approach to spreading activation.

5. PROPOSED SEMANTIC INFORMAITON RETRIEVAL

The figure illustrates the overview of the system.

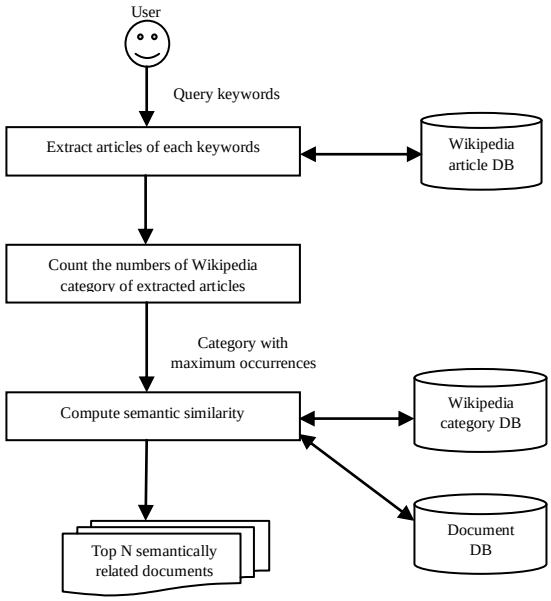

Figure. 1 Overview of proposed system

The system intends to utilize the wide range of knowledge from Wikipedia. The system uses the method of spreading activation for computing semantic similarity using category tree of Wikipedia. It can reduce the effort of building ontology for computing semantic similarity. It will produce the semantically related results.

The steps of the system are as follows. When the user enters the query as keywords he wants to search, the system will extract the corresponding Wikipedia articles of each keyword in the query. Then system will collect the lists of the categories of each article and count the categories which belong to the articles. The system will search the required information based on the category having the most occurrences. To rank the information according to their semantic similarity, the system will compute semantic similarity using spreading activation strategy based on the category tree of Wikipedia. So, the system has two main parts: one is searching for the category that has the most occurrences and another is computing the semantic similarity able to retrieve the semantically relevant information.

6. COMPUTING SEMANTIC SIMILARITY WITH SPREADING ACTIVATION STRATEGY

To compute semantic similarity for our IR system, firstly we extract the Wikipedia categories of each query key word. Then we also extract the Wikipedia categories of document title in the document database. Before we do the latter extraction, we need to search for the corresponding Wikipedia articles of the document title. Finally, we use all these categories extracted as the nodes of the category tree of Wikipedia and apply the spreading activation method to this category tree to get semantic similarity value.

The followings are the node input function, output function and semantic similarity computing function.

$$I_j = \sum_i O_i \quad (1)$$

$$O_j = \frac{A_j}{D_j * k} \quad (2)$$

(3)

$$\text{Similarity Value} = \frac{\sum_{\forall A_i \in Act} A_i}{|Act| * \max(A_i)}$$

Where the variables are defined as:

O_i : Output of node i connected to node j

A_j : Activation value of node j

k : iteration number

D_j : Out degree of node j

I_j : input to node j from the child node i (is also Activation value of node j)

Act : set of activation value

The activation process is iterative. All the original nodes take their occurrences as their initial activation value. And the activation values of all the other nodes are initialized to zero. Every node propagates it activation to its parents. The propagated value (O_j) is a function of its activation level. After a certain number of iterations, the highest activation value among the nodes that are associated with each of the original node is retrieved into a set Act = $\{A_1, A_2, \ldots, A_{n+m}\}$. Then the similarity value is computed using the values from the Act set with the equation (3). The similarity value is normalized to value between 0 and 1.

7. CONCLUSION

In this system, we proposed the use of Wikipedia category tree and spreading activation strategy to compute semantic similarity. This system uses Wikipedia as an ontology. So it can reduce the effort of expert required to build ontology. Spreading activation strategy has produced excellent results for other semantic related system such as word sense disambiguation, semantic similarity computing using ontologies and describing documents. Therefore, the proposed system uses this method in the information retrieval system to produce the semantically related information along with the information required for the user.

8. REFERENCES

[1] Gabrilovich, E. and Markovitch, S., "Computing semantic relatedness using wikipedia-based explicit semantic analysis," Proc. Of the 20th International joint Conference on Artificial Intelligence (IJCAI'07), pp. 6-12.

[2] Sapkota, K. Thapa, L. and Pandey, S., "Efficient information retrieval using measures of semantic similarity," Nepal Engineering College.

[3] Yazdani, M. and Popescu-Belis, A., "A random walk framework to compute textual semantic similarity: a unified model for three benchmark tasks".

[4] Zesch, T. Müller, C. and Gurevych, I., "Using wiktionary for computing semantic relatedness," Proc. Of the Twenty-Third AAAI Conference on Artificial Intelligence (2008), pp. 861-866.

[5] Syed, Z. S. Finin, T. and Joshi, A., "Wikipedia as an ontology for describing documents," Association for the Advancement of Artificial Intelligence, 2008, pp. 136-144.

[6] Zhiqiang, L. Werimin, S. and Zhenhua, Y., "Measuring semantic similarity between words using wikipedia," International Conference on Web Information Systems and Mining, 2009, pp. 251-254.

[7] Thiagarajan, R. Manjunath, G. and Stumptner, M., "Computing semantic similarity using ontologies," the International Semantic Web Conference (ISWC), 2008, Karlsruhe, Germany.

[8] Hajian, B. and White, T., " Measuring semantic smilarity using a multi-tree model," 2011.

[9] Milne, D. and Witten, I. H., "An effective, low-cost measure of semantic relatedness obtained from wikipedia links," Proc. Of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA, pp. 25-30.

[10] Yeh, E. Ramage, D. Manning, C. D. Agirre, E. and Soroa, A., "WikiWalk: Random walks on wikipedia for semantic relatedness," Proc. Of the 2009 Workshop on Graph-based Methods for Natural Language Processing, ACL-IJCNLP 2009,Suntec, Singapore, Aug 7. 2009, pp. 41-49.

[11] Strube, M. and Ponzetto, S. P., "WikiRelate! Computing semantic relatedness using wikipedia," Proc. Of the National Conference on Artificial Intellignece, 2006, volume 21.

[12] Gouws, S. Rooyen, G. and Engelbrecht, H. A., "Measuring conceptual similarity by spreading activation over wikipedia's hyperlink structure," Proc. Of the 2nd Workshop on " Collaboratively Constructed Semantic Resources", Coling 2010, Beijing, August 2010, pp. 46-54.