Comparative Analysis of Various Approaches for Semantic Information Retrieval

Poonam Chahal
Research Scholar,
YMCAUST, Faridabad
Poonmandal.fet@mriu.edu.in

Manjeet Singh
YMCAUST, Faridabad
mstome2000@yahoo.com

Suresh Kumar
FET, MRIU, Faridabad
enthusk@yahoo.com

Abstract*: Semantic similarity between words/sentences/documents has been studied extensively in recent years due to exponentially increase in World Wide Web. It has become necessary to minimize the problem of information retrieval faced by the users of internet to understand the semantic form of information that is presented on web. But, the present retrieval systems consider only the syntactic structure of the information so the methods or approaches which can help us to understand the semantics of the Information presented in any form is gaining importance. Researchers have considered semantic form of contents and proposed or implemented different ways of finding the relevant information from the web as per the user's expectation. In this paper we have tried to discuss the approaches given by different researchers in the area of semantic similarity between the words/sentences/ documents present in WWW and then tried to conclude their salient features and limitations.*

Keywords: *Semantic Web, Ranking, Parsing, Syntactic, Lexical, Semantic, Similarity.*

I. INTRODUCTION

The WWW (World Wide Web) is the large information resource in which information is present in the form of web pages that are linked to each other [1]. Due to information overload the information retrieval by the user's of WWW should be performed in effective and efficient way so that the relevant information can be retrieved by the user as per their expectations [2]. For the minimization of the problem of information retrieval due to information overload on WWW it has been necessary to find effective and efficient methods to understand the semantics of the information that is presented on web.

Many researchers in this domain have tried to explore different methods for finding the semantic similarity between the words/sentences/documents. Ontology is used for understanding the semantics of any information. Ontology is the conceptual description of the information contained in the text [10]. This conceptual description can be done using the graphical structure in which nodes represent the words/concepts and the edges between any two nodes represent the relationships that exists between these words/concepts. In the section II we briefly describe ontology and how it helps for semantic information, then in section III we summarize the approaches, finally in section IV the new proposal and conclusion is given.

II. LITERATURE REVIEW

Cordi V et. al. [3] has given an ontology-based similarity between sets of concepts the authors evaluate the information presented in one document with another document with respect to the ontology. In their approach the authors extracted the concepts of the documents and computed the semantic similarity between them with respect to a given ontology using by modifying the dijikstra algorithm of the graph theory. In this approach the authors only considered the set of concepts rather than the relationship and the type/kind of relationship between them.

Pisharody A et. al. [8] proposed a search engine technique using keywords relations. In this paper the drawback of keyword based approach is overcome by creating a database that consists of words and their relations in addition to keywords. The LGP parser is used to parse the web pages. From each line that is

present in the web pages the noun, adjective, verb, determiner, preposition, etc. are identified. Out of these, the noun, adjective and verb are stored in the database. The process of normalization is used to remove duplicate values. Each word extracted above is fed into WordNet to determine the sets of relations. Thus the database is constructed having words and its relations. When the query is given by the user, it is parsed by retrieving the noun, adjective and verb. The query word is searched in the database created for the webpage and all its relations are retrieved. If the word is not present in the database then reverse lookup algorithm is executed in which instead of searching the word, the relation part is searched. Although the authors have tried to remove the keyword based similarity but then even the similarity results are not upto the user expectations as they do not consider each and every concept and also all the relations that exists between them.

Thiagarajan R. et. al [10] proposed computing semantic similarity using ontology's. In this paper authors represented the web page can be represented either Bag of Concepts (BOC). In BOC the concepts are taken from the web page to represents the web page more semantically. Now for computation of semantic similarity between the web pages the authors used process of spreading, which means including additional related terms to an entity by referring to ontology such as Word Net, Wikipedia. For the spreading process two schemes are used one is set spreading and other is semantic network.

Then the similarity computation is computed by cosine similarity.

$$sim_{cos}\ (e_j, e_k) = \frac{V\ (e_j).\ V\ (e_k)}{|V\ (e_j)\ |V\ (e_k)|}$$

V. Oleshchuk [7] discussed ontology based semantic similarity comparison of documents to reflect the semantic relationship between concepts. The authors performed the document articulation rather than comparing the raw content of the documents. The process of document articulation done with respect to the given ontology produces a graph that is obtained with the help of the ontology. Similarly each document can be articulated and the graph obtained is the document ontology. Now, for the comparison of the text of two documents, the document ontology obtained by articulation of these documents with the help of a given ontology are labeled level-wise. In Fig. 1 and Fig. 2, the example showing the two document ontology for the domain transportation and the ontology taken for obtaining the document ontology is given which are used by the authors to show their results. Thus, the similarity between two sub-ontology can be obtained at each level which is represented as a vector that have values 0 or 1. If the node of the two sub-ontology is same at any level then it is assigned score 1 otherwise 0.

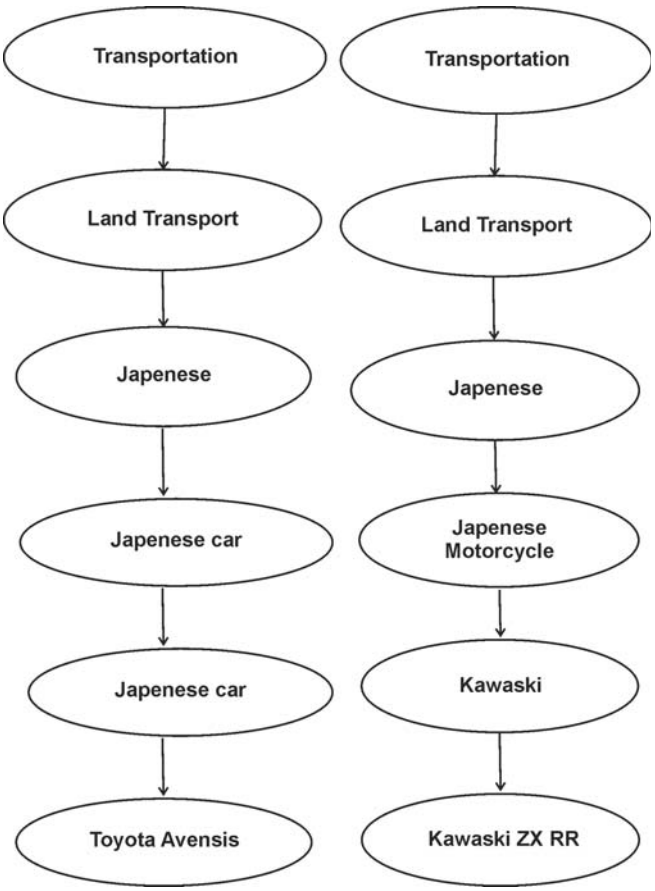

Fig 1: Ontology O1 and O2 for text t1 and t2

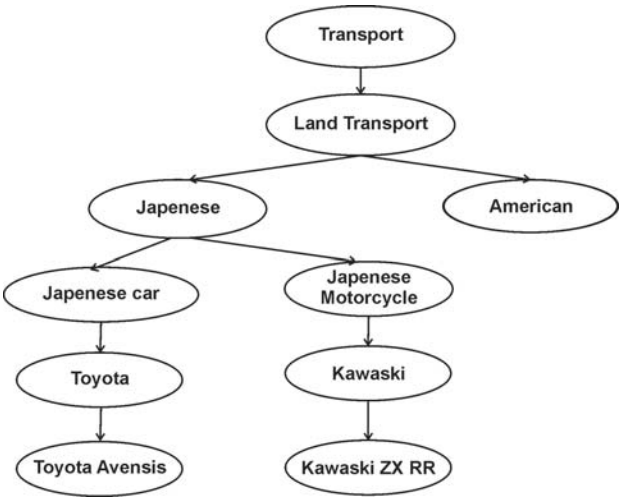

Fig 2: Ontology for Transportation

In the discussion given in [7] the authors only considered the concepts at each level and did not discussed about the edges which represents the relationship between the concepts. To find the true semantics it is necessary to use and understand the relationships between the concepts.

B. Hajian et. al. [4] proposed a method of measuring semantic similarity using a multi-tree model In this paper the authors proposed the new method for semantic similarity based on the knowledge that is extracted from ontology and taxonomy. The technique described uses multi-tree similarity algorithm to measure similarity of two multi-tree constructed from taxonomic relations between entities in ontology. In Fig. 3 and Fig. 4 the multi-tree obtained for the two documents taken by the authors [4] and the combined tree obtained for the two documents taken is shown. The similarity comparison is done by comparing the feature list representing the concept, i.e. each concept in the approach given by [4] is represented by the features describing its properties. Though the authors considered the edges representing the relationship between the concepts but they did not considered the number and the type of relationship which can exist between two concepts in the ontology or the document.

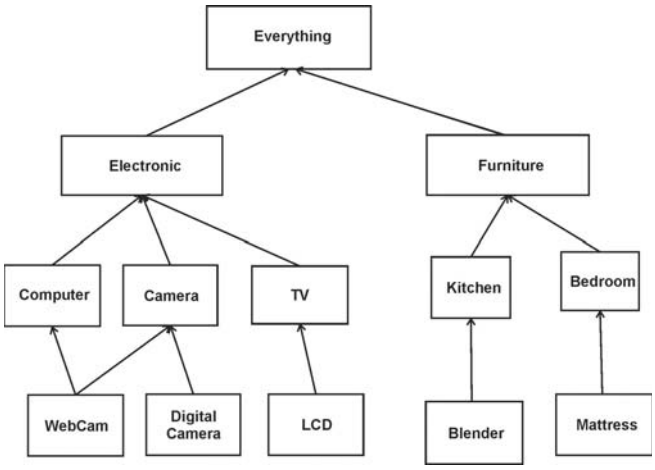

Fig 3(a): Multi-tree representing transaction d1.

Li Y. et. al. [6] proposed relation based search Engine. In this paper the authors proposed a semantic search engine ONTOLOOK which considers relations between the concepts. A page will be returned to user only when it includes the relationship between keywords. First thing the ONTOLOOK does is analyzing keywords input by the user. Then, the keywords will be assembled to some concepts pairs and these pairs are sent to the ontology

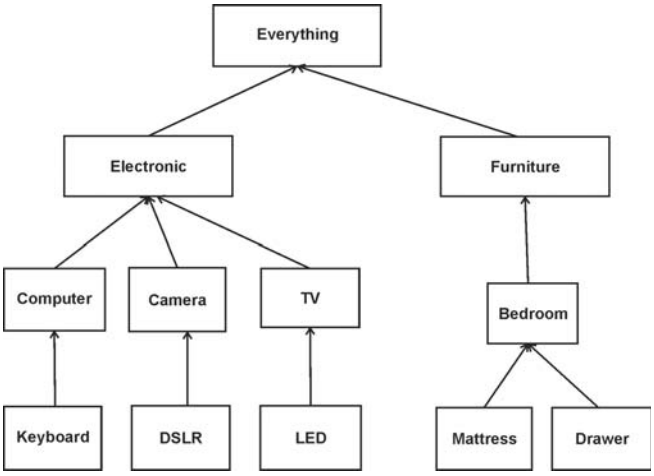

Fig 3(b): Multi-tree representing transaction d2.

database to retrieve all the relations defined by the ontology between concept pairs. After all relations retrieved the concept-relation graph is formed based on these relations and concepts. Then Ontolook will cut some arcs from the graph and construct subgraphs. Finally the system fetches the relation and corresponding keyword pairs from each arc in subgraphs to form property-keyword candidate set and then it is sent to the database to get a retrieved result set for the user. The ranking algorithm used by the authors is same as of Google Page Rank. So for further improvement of the ranking of the result-set this has to be modified in an efficient manner to consider the relation that is the most important part to understand the semantics of the document.

Lamberti F. et. al. [5] used relation-based page rank algorithm for semantic web search engines to focus on the concepts that exist within the document and also on the relationship that exists between those concepts which an extension given to the [6]. The authors fully explored the number of relationships that exists between the concepts in a document with respect to the number of relations that are presented in the given ontology between same concepts. They proposed a technique to exploit the relevance feedback and post process result-set to develop a ranking strategy which considers relation between keywords which is given in web page. A page rank algorithm which is based on relations that exists between the concepts is given which can be used in conjunction with the semantic web search engine. The approach is to construct the graph of underlying ontology, query, page annotation and page sub graph. Then they computed probability for a page to be selected by taking the factors of number of relation in ontology, query and page annotation and sub graph. This considers the ontology graph, query graph, and annotation of page and its sub

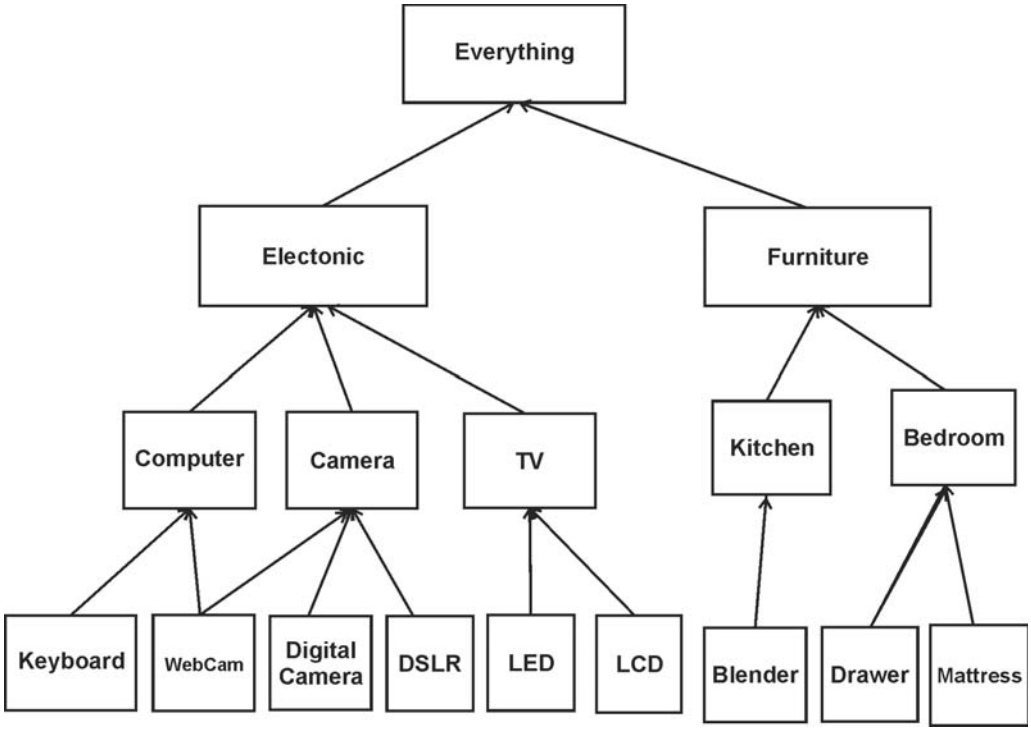

Fig 4: A multi-tree combined for previous multi-tree

graph. Now, some concepts can be possible which are not related to any other concept in the annotations but that can be of user interest. So, the probability that each concept is related to other concept is modeled using graph theory i.e each concept related to at least another concept in the query is equivalent to considering all possible spanning trees.

Giannis V. et. al. [11] proposed semantic similarity methods in wordnet and their application to information retrieval on the web In this paper the authors investigated the approach to compute semantic similarity by mapping terms (concepts) to an ontology and by examining their relationship in that ontology. The proposed method is capable of detecting semantic similarity between documents which are not lexicographically similar terms. In first part the authors proposed discovery semantically similar terms using wordnet and in the second part proposed SSRM method to generate semantic similarity.

In this approach only the query terms are expanded and reweighting. The document terms dj are computed as $d_j = tf * idf$, it means they are neither expanded nor reweighted.

III. COMPARATIVE ANALYSIS

In the work surveyed the main focus is on introducing the semantics either by taking ontology or relationship that exists between the concepts. It has been necessary to compare the semantic similarity between the documents to find the true value /relevance of similarity between the documents.

Some researchers have used the approach of extracting keywords from the documents by just keeping the noun, verb and adjective present in the document. These keywords are then stored in a database which contains words along with relations using WordNet. The query words are then searched in table and also the corresponding relations. In this way the authors [8] tried to search the words of the query in the document and also the related words. These documents can be parsed and the keywords extracted can be extended using WordNet and then making a tree of other document and then trying to merge the two graphs using ontology [4][9]. This will give the hierarchical representation of the documents in the form of ontology but the concepts relations and their types along with the relevance are ignored.

To make the system effective the interface for the search engine like ONTOLOOK in which the user is allowed to give the words along with relation that is in the mind of the user [6]. Then, the ranking of the documents can be done using PageRank algorithm. Another author [5] gave the approach of the relation based algorithm for ranking the documents as extension to the ONTOLOOK [6]. In this the relationship between the words are considered in the document, query and also the ontology. Although, the interface for taking the user query in the form that the idea

which is present in the mind of the user is taken, but then also the ranking of the documents is still done using the traditional approach.

Computation of semantic similarity can be done using ontology [10]. In this method the documents are analyzed as bag of concepts and they are extended using WordNet, then the similarity is calculated. In this the authors considered words and their relation but not the type of relation.

Similar approach for finding the similarity comparison of documents using ontology [10] is used by representing the document as sub ontology obtained by the process of document articulation. Then the subontology obtained for the documents is compared level wise to get the vector value of their similarity. The comparison between various approaches is shown in Table 1.

Table 1: Comparative Analysis of Various Approaches for Finding Similarity

Author	Concepts	Relations	Ontology
Cordi	√		√
Pisharody	√	√	
Thiagarajan	√	√	
Oleshchuk	√		√
Hajjan		√	√
Li		√	√
Proposed	√	√	√

IV. CONCLUSION AND FUTURE SCOPE

Some researchers have used the concept of assigning weights to the terms present in the documents with respect to the query words [11]. Also the query words can be expanded by adding the synonym, hypernym and hyponym and these words are also weighted and then the similarity can be found between the query and the documents for the relevant ranking of the documents.

Different approaches have been followed by different researchers in the field of semantic similarity. Some have taken only keywords, others have also considered relationship that exists between these keywords. These relationships are taken with the help of WordNet, ontology etc.

We will try to find the semantic similarity between the sets of documents. In our approach we will try to consider keywords along with their relationship types and their weights using ontology. One way of calculating semantic similarity is by using ontology and then applying some NLP techniques to find similarity e.g. calculating syntactic similarity.

REFERENCES

[1]. Berners-Lee T., Hendler J., and O. Lassila, "The Semantic Web," Scientific Am., 2001.

[2]. Brin S. and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Proc. Of 7th Int'l Conf. on World Wide Web(WWW '98), pp. 107-117, 1998.

[3]. Cordi V, Lombardi M, Viviana M, "An ontology-based similarity between sets of concepts", WOA, 2005.

[4]. Hajian B., and Tony W., "A method of measuring semantic similarity using a multi-tree model proceedings IJCAI 2011 - 9th Workshop on intelligent techniques for web personalization & recommender systems (ITWP'11) Barcelona, Spain, 16 JULY 2011.

[5]. Lamberti F., Sanna A., and C. Demartini, "A relation-based Page Rank algorithm for semantic web search engines", IEEE Trans Knowledge and Data Eng., vol. 21, no. 1, Jan 2009.

[6]. Li Y., Wang Y., and X. Huang, "A Relation-Based Search Engine in Semantic Web", IEEE Trans. Knowledge and Data Eng., vol. 19, no. 2, pp. 273-282, Feb. 2007.

[7]. Oleshchuk V., and Asle P., "Ontology Based Semantic Similarity Comparison of Documents", Proc. of IEEE 14th workshop on database and expert systems applications, 2003.

[8]. Pisharody A. and H.E. Michel, "Search Engine Technique Using Keyword Relations", Proc. of Int'l Conf. on Artificial Intelligence (ICAI '05), pp. 300-306, 2005.

[9]. Takale S., and Sushma N., "Measuring Semantic Similarity between Words Using Web Documents", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 1, No.4 October, 2010.

[10]. Thiagarajan R., Manjunath G., and Markus S., "Computing semantic similarity using ontologies" ISWC 08, the International Semantic Web Conference (ISWC), Karlsruhe, Germany, 2008.

[11]. Varelas G., Voutsakis E., and Paraskevi R., "Semantic similarity methods in Wordnet and their applications to information retrieval on the web", WIDM ACM Transactions, Bermen , Germany, 2005.

❐