International Journal of Current Engineering and Technology
ISSN 2277 - 4106
© 2013 INPRESSCO. All Rights Reserved.
Available at http://inpressco.com/category/ijcet

Research Article

An Ontology Based Approach for Finding Semantic Similarity between Web Documents

Poonam Chahal [A]*, Manjeet Singh [A], Suresh Kumar [B]

[A] YMCAUST, Faridabad, India
[B] FET, MRIU, Faridabad, India

Accepted 25 November 2013, Available online 01 December 2013, **Vol.3, No.5 (December 2013)**

Abstract

Recently, many semantic web search engines have been developed like Ontolook, Swoogle, etc which help in searching meaningful documents presented on semantic web. In contrast to this the commonly used approach is based on matching keywords extracted from the document which is known as lexical matching. But there exist the documents that contains same information but using different words i.e. one document using a word and other document using synonym of that word. So, when similarity of such documents is computed through lexical matching it will not give true results of similarity computation. In this paper we have proposed a semantic web document similarity scheme that relies not only on the keywords but on conceptual instances present between the keywords and also considers the relationships that exists between the concepts present in the web pages. We explore all relevant relations between the keywords exploring the user's intention and then calculate the fraction of these relations on each web page to determine their relevance and similarity with the other documents. We have found that this semantic similarity scheme gives better results than those by the prevailing methods.

***Keywords**: Semantic Web, Ontolook, Swoogle, Indexing, Semantic, Similarity.*

1. Introduction

The World Wide Web (WWW) is large information resource centre in which information present in the form of web pages is interlinked with each other. With the enormous amount of information presented on the web, it has been difficult to find or access relevant information by the wide categories of users of web and present or maintain the information by any machine. This is because web content is presented primarily in natural language, and targeted to human reader. However, some information retrieval tools, such as Google, Yahoo etc. are being used by human reader in order to access the desired information.

A search engine is a program that searches for information stored on WWW. The search engine works for abstraction and identification of information stored in WWW by using a spider, robot or crawler to fetch the documents as much as possible to achieve its goal. Another program, called indexer, then process these documents and creates an index of these documents depending information contained in them. Every search engine uses its own proprietary algorithm to create its indices such that only meaningful results are returned for each user query. But, the result-set produced by the search engine are not up to the user expectation as there is a wide gap between the techniques required for automatic processing of information by the search engine to produce meaningful results/information and the techniques being used at present to process information presented in web document designed mainly for human readability. The next generation of search engines must address this problem and deal with it in a layered architecture of semantic web to overcome this limitation.

The semantic web visualized by Tim Berners-Lee is a collection of resources and their description thereby allowing machines to interpret data/description in order to maintain/organize the resource for information processed by computer program or by any service. Recently, many semantic web search engines have been developed like Ontolook, Swoogle, etc which help in searching meaningful documents presented on semantic web.

2. Related Work

In fact, the process of retrieving relevant information with the help of a search engine is very crucial. The indexing of documents by the search engine can be done only by finding the similarity between the fetched web pages. Similarly, the ranking by a search engine is done by finding similarity between the query given by the user and the web page. Some attempts have been made in finding

*Corresponding author: **Poonam Chahal** is a Research Scholar.

the similarity between the documents but still the results provided by similarity detection techniques are not up to the user's expectations.

In General, the similarities between the documents and the knowledge that different documents have similarity is of great importance from several aspects such as they relate to same fields or concept or interest and also many applications like removing or identifying duplicate pages while crawling, indexing, ranking process to provide user relevant and meaningful result-set, finding related documents which are on similar or same topics to know different versions of the documents detecting plagiarism, multi-document summarization, etc.

To incorporate the semantic aspect in the search engine requires its development in the form of layered architecture to handle semantic web focuses on considering the concepts and relations between the concepts that exist in the document. In contrast to this the commonly used approach is based on matching keywords extracted from the document which is known as lexical matching. But there exist the documents that contains same information but using different words i.e. one document using a word and other document using synonym of that word. So, when similarity of such documents is computed through lexical matching it does not give true results of similarity computation.

There are various techniques based on Natural Language Processing, Lexical analysis, Semantic analysis, Ontology based matching etc. for computing the similarity of the documents. Using NLP techniques in document processing we can obtain the selected informative words and then visualization of documents is done by disambiguation of those words that have several meaning.
In Lexical matching approach only keywords are taken into consideration. The vector space model (VSM) is used in lexical matching i.e. for each document the vector space model is constructed consisting of the keywords extracted from the document and after that the similarity can be computed using Cosine similarity, Jaccard similarity, Dice similarity, etc. In Semantic Analysis when the documents are analyzed semantically and then their similarity is computed by taking not only keywords but also the concepts, synonyms of the words and relation between the concepts. The similarity can be represented using graph theory, relational algebra. In ontology based matching the similarity computation between the documents is done by using the ontology like Protege, Sweet, WordNet etc. The documents concepts can be extracted and extended using the ontology in which the concepts extracted are extended with the hyponym, meronym, synonym etc. The parameters associated with the ontology taxonomic hierarchy can be length of shortest path, depth of most specific common subsumer, density of concepts of the shortest path, density of the concepts from the root to the most specific common subsume.

Fabrizio L. *et. al.* proposed a Relation Based Page Rank algorithm for Semantic Web search Engine. In this paper authors proposed a technique to exploit the relevance feedback and post process result-set to develop a ranking strategy which considers relations between keywords which are given in a web page. The algorithm relies on the information that is to be extracted from user queries and the resources with annotation like web pages. The page relevance is calculated using probability that page actually contains relation whose existence was assumed by user at time of query definition.

Vladimir O. *et. al.* have given Ontology Based Semantic Similarity Comparison of Documents. In this work the authors considered ontologies as knowledge structures that specify terms, their properties and relations among them to enable knowledge extraction from texts. They represented ontologies using a graph-based model that reflect semantic relationship between concepts and apply them to text analysis and comparison. Instead of raw document comparison they compared document footprint enhanced with concepts from the ontology (using different enhancement algorithms). The result of this process may be that documents which appear to be not similar prior to the enhancement may become similar (semantically on some abstraction level) after the enhancement using ontology. The authors have given the ontology extraction algorithm and similarity between sub-ontologies.

B. Hajian *et. al* have given a method of measuring semantic similarity using a multi-tree model. In this paper the authors proposed the new method for determining semantic similarity based on structure- knowledge extracted from ontology and taxonomy. The technique described by them uses multi-tree similarity algorithm to measure similarity of two multi-tree constructed from taxonomic relations between entities in ontology. Another multi-tree is built from the two trees obtained from each document. The similarity between two concepts is measured by commonality of their features. Each concept is represented by feature describing its properties; a similarity comparison involves comparing the feature list representing the concept. The similarity between two documents is equal to the value of similarity of the root node.

A. Pisharody *et. al.* proposed a search engine technique using keywords relations. In this paper the drawback of keyword based approach is removed by creating intelligent database that consist of words-relations in addition to keywords. In this approach the web pages are parsed using LGP Parser. Each line in the web pages contains noun, adjective, verb, determiner, preposition, etc. Out of these words the noun, adjective and verb are stored in the database. The duplicate values are removed by normalization. After this, each word is fed into WordNet to determine the sets of relations. Thus the database has words and their relations. When user gives the query and it is parsed retrieving the noun, adjective and verb and then the word is searched in corresponding database of the webpage and retrieve all its relation. If the word is not present in the database then reverse lookup algorithm is executed in which rather than searching the word, the relation part is searched.

R. Thiagarajan *et. al* proposed computing semantic similarity using ontology's. In this paper the authors have given that the web page is represented either by Bag of words(BOW) or Bag of Concepts(BOC). In BOW

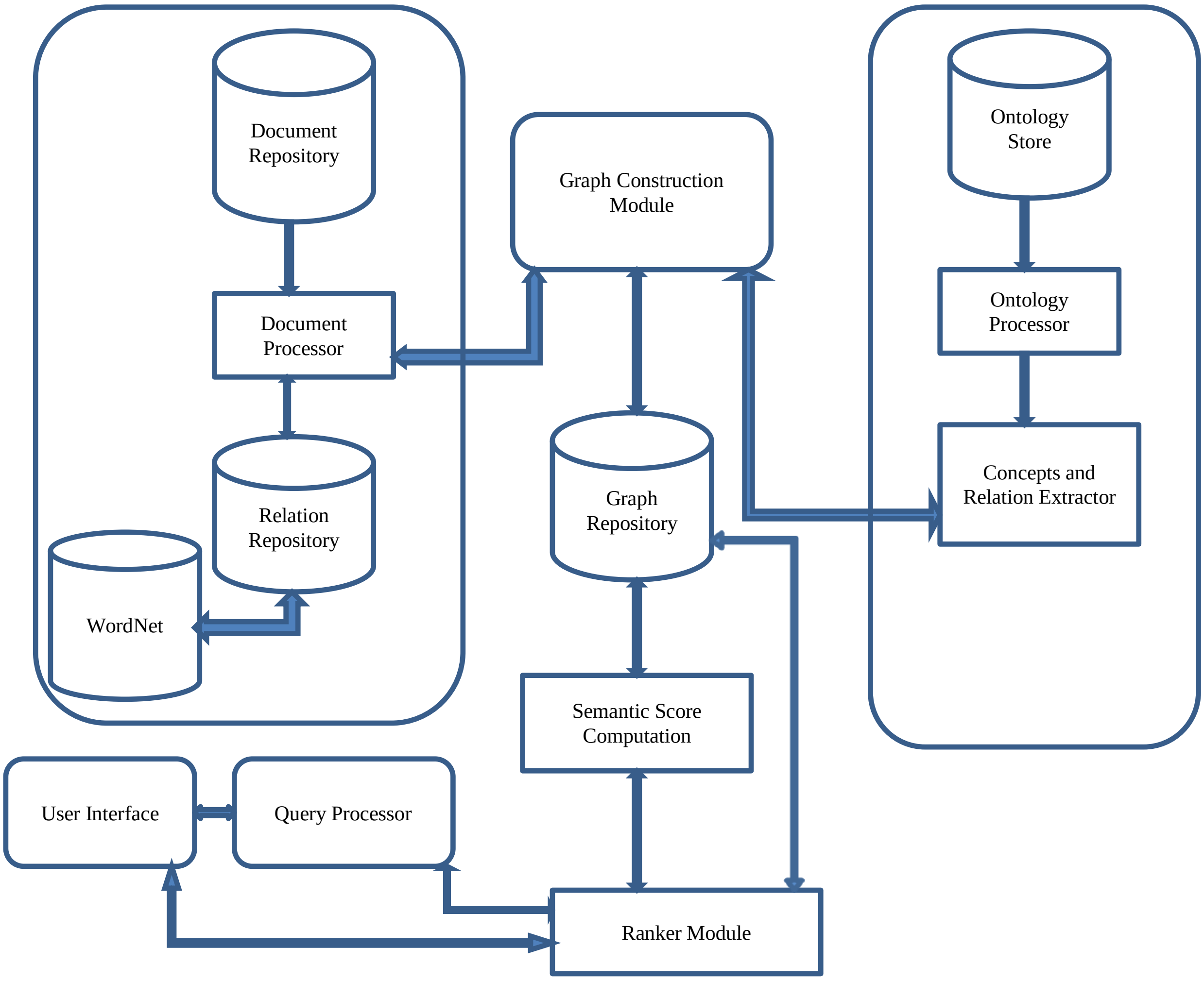

Figure 1: Architecture and System Flow Diagram of proposed Semantic Similarity Model.

approach, only keywords are taken so it lacks intelligence while in BOC the concepts are taken from the web page so it represents the web page more semantically. Now to compute semantic similarity between the web pages, the authors used the concept of spreading which is the process of including additional related term to an entity by referring to ontology such as Word Net, Wikipedia. For spreading two schemes are used one is set spreading and the other is semantic network. Then the similarity computation is computed by cosine similarity.

Hung C. *et. al.* proposed a New Suffix Tree Similarity Measure for Document Clustering. In this paper the authors proposed a new similarity measure to compute pair wise text-similarity based on suffix tree document model. Then the similarity is applied in group agglomerative hierarchical clustering and a new suffix tree document clustering algorithm is developed. The framework of this data model is the document representation as a feature vector of words that appear in document. The term weights are also contained in each feature vector. Similarity is calculated using Cosine, Jaccard, Euclidean distance measure.

Fernando S. *et. al.*presented a semantic similarity approach to paraphrase detection. In this the authors used the approach using similarity matrix for paraphrase identification. The authors represented each sentence by a binary vector (with elements equal to 1 if a word is present and 0 otherwise), a and b. The similarity between these sentences can be computed using the following formula:

$Sim(a,b)=aWb/|a||b|$

where W is a semantic similarity matrix containing information about the similarity of words.

In all these contributions the main focus is on introducing the semantics either by taking ontology or relationship that exists between the concepts. The researchers either tried to use the chunk based approach, adding semantics by extending the keywords using WordNet, or by matching the string and then adding semantics. This makes it necessary to compare the complete semantic similarity between the documents to find the true value of similarity between the documents.

3. Proposed Semantic Similarity for Semantic web Documents

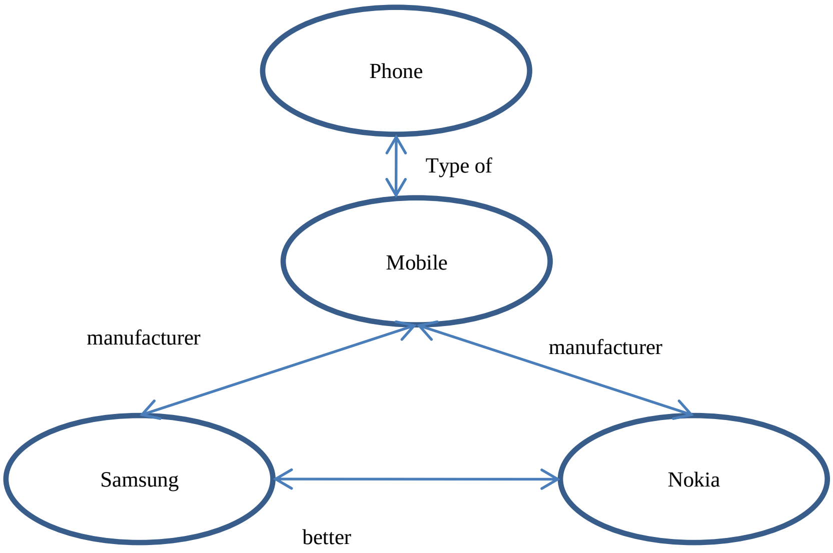

Figure 2a: graph of document A.

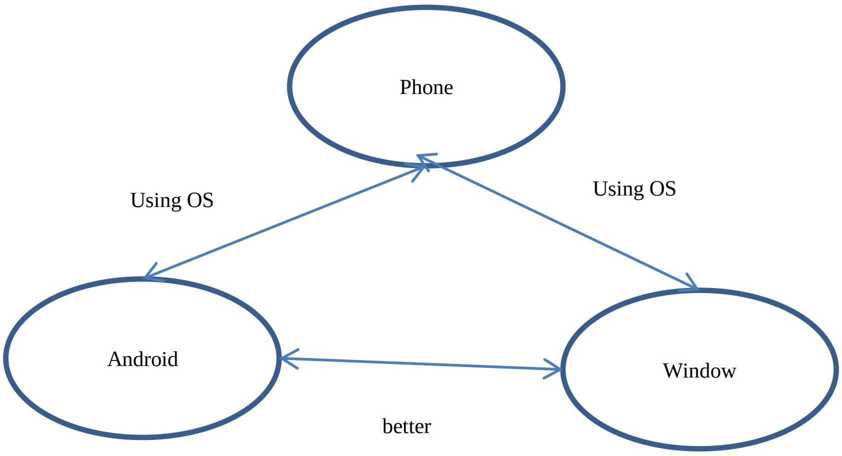

Figure 2b: graph of document B

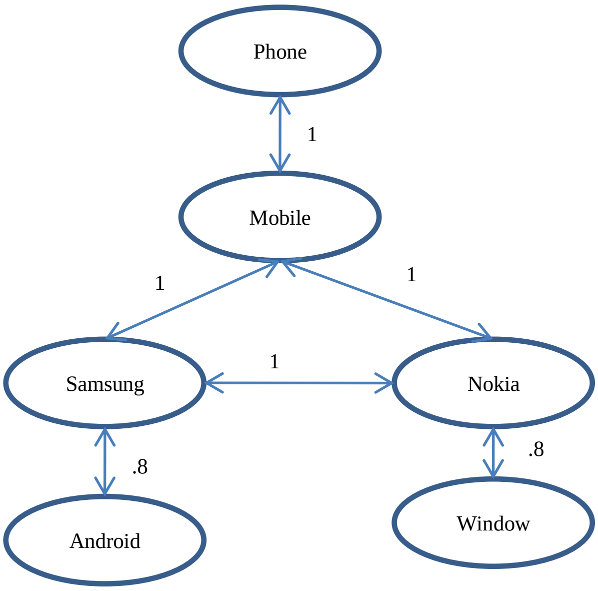

Figure 3a: Graph of Document B after spreading.

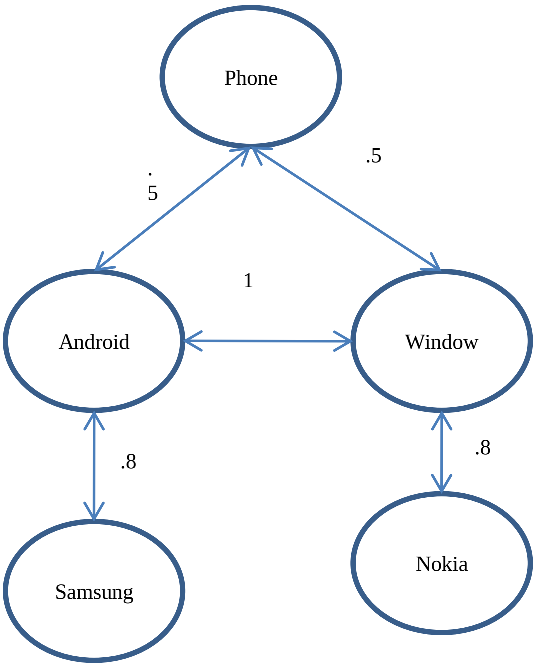

Figure 3b: Graph of Document B after spreading.

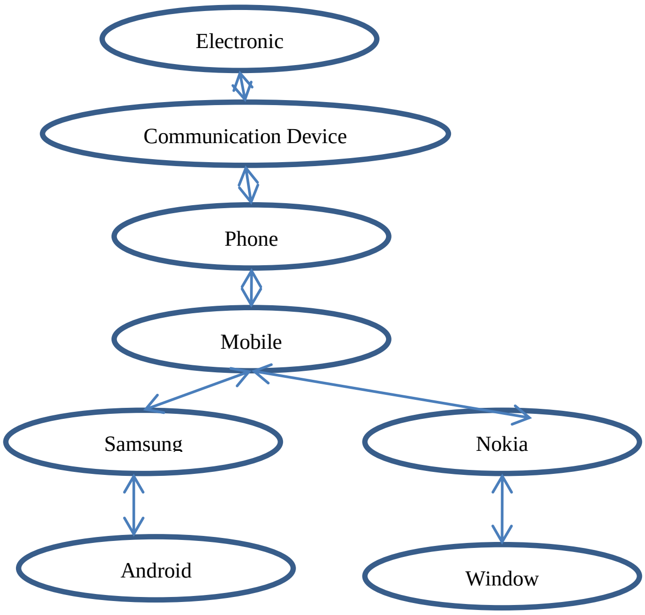

Figure 4: Given ontology O.

The semantic similarity between the query and documents for ranking can be done by keeping the user view in mind i.e. by considering the query and extending the query using ontology. There can also be possibility of constructing ontology of a document.

The document can be parsed and the keywords extracted can be extended using WordNet and then making a tree of one document and similarly making a tree of other document and then trying to merge the two graph using ontology.

Some researchers have used the approach of extracting keywords from the document and just keeping the noun, verb and adjective and removing rest of the keywords. Then storing them in a database and comparing the list using ontology.

In this paper we are giving an ontology based approach for finding the semantic similarity between the semantic web documents. The overall system architecture is given in Figure 1. The main components of the architecture of the system are ontology processor, graph construction module, ranker module and document processor. In this approach first the document processing is done by extracting the keywords using syntactic analysis and making a VSM for the document representing the terms along with the frequency. Now, to extract the relation from the document we uses a relational repository in which the types of relation exists along with the weightage assigned according to fuzzy set theory and description related to that relation is also present. Now, by considering the relational repository for document processing we can retrieve the concepts relationship that exists in the documents. Then, spreading of the document is done by using the given ontology in which all the concepts and relationship among the concepts is present. After the spreading process the graph for each document can be constructed in which the nodes represent the concept and the edges represent the relationship between the concepts. This graph construction is done for each document by considering a document dictionary in which terms along with the synonyms are present to consider all the words synonyms. The documents graph can now be used to find semantic similarity between the documents by considering the similarity between not only the nodes but also the edges between the nodes which represents the relationship between the nodes thereby considering complete semantics between the documents. Thus, the similarity between the two graphs of the documents is calculated using the probability

$P(A\cap B)=1-(n(G(A\cap B))+r(G(A\cap B))/(n(G(A))+n(G(B))+r(G(A))+r(G(B))$

where $n(G(A\cap B))$ and $r(G(A\cap B))$ represent respectively the number of nodes and numbers of relations that are common in both the graphs of the documents for which we want to find similarity. The $n(G(A))$, $n(G(B))$ represents the number of nodes in the graph of the two documents A and B. Similarly $r(G(A))$ and $r(G(B))$ represents the relationship that exists between the nodes in the respective graphs of two documents. Figure 2 represents the graph of two documents. In this we have assumed that document A is having content as={android based phones are better than window based phone} and document B is having content as={Samsung based mobiles are better than nokia based mobiles}. In Figure 3 we represent the extended graph of the documents by using the process of spreading in which we consider the relation Table 1, document dictionary and the ontology which is shown in Figure 4. And finally we calculated the similarity by finding the value of by taking $n(G(A\cap B))=1$ and $r(G(A\cap B))=2$.

The corresponding value obtained from figure 3 graphs $n(G(A))=6$, $n(G(B)) =5$ and $r(G(A))=6$, $r(G(B))=5$. So $P(A\cap B)=.86$. Similarly we have taken more than 50 examples of the documents containing the content of same type and representing the content by different keywords i.e. the idea that is to be conveyed by the documents is same but it is given in different way. By our approach we have tried to capture the view of the user in which the intension of what user wants to retrieve is taken into account. The similarity of these documents cannot be computed using the lexical matching approach. In fact, the idea of the document is same but the representation of idea is different. But keywords based approach takes into account only the words. Thus the similarity found for such documents is not up to the mark. But our approach not only takes the keywords but also the relationship that exists between the keywords.

Table 1: Relation Table having Weights along with the description

SNO	Relation	Weights	Description
1	Type of	1	--------
2	Is a	1	--------
3	of	.8	--------
4	Part of	1	--------
5	Kind of	1	--------
6	using	.5	--------
7	At	1	--------
8	Has	.9	--------
9	Through	.9	--------

In this paper an ontology based approach for finding the semantic similarity is given which not only considers the keywords but also the relationship between the keywords to find the true value of similarity between the documents. In future we will also try to find the similarity by using not only the single ontology but any type of ontology that can be built and used.

4. Performance Analysis

Performance of our approach for finding semantic similarity between the semantic web documents definitely depends on how keywords and associated concept relations are extracted from the document, then on the process of spreading used to create and would depend from domain to domain as well as formulation of concept relations. We have compared the performance of our semantic similarity scheme with the similarity computed using keyword based approach. From these pages we determined the actual similarity of the pages using the keyword based approach and also with the novel approach given in this paper. And in maximum number of cases we found our approach giving better similarity measurement. The results obtained from the novel approach and have been presented in Table 2.

Table 2: Results of similarity for keyword based approach and proposed semantic similarity approach.

S.No	Document	Document	Keyword	Novel
1.	Maintenance	Maintenance	.67	.78
2.	Market	Market	.75	.57
3.	I like	Teaching is	.5	.72
4.	Android	Window	0	.64
5.	Blue-Ray	DVD player	.4	.67
6.	Apple	Dell Laptop	.5	.4

For deep analysis of the performance of our approach with lexical matching, we further looked to the pages retrieved from Google search engine having similar content but not represented with same words. The large number of PDF files giving similar content was taken and the summarization process applied to the documents and then the keywords extracted from the documents summarization and also along with their weightage. The terms which were extracted with similarity greater than the threshold value were taken into account for the process of spreading with the help of given ontology. The graph constructed from spreading process was then scanned to get the number of nodes along with the concepts representing the nodes. Then the similarity between the concepts and the relationship that exists between the nodes was then calculated using the approach given in this paper. We found that for maximum number of documents the approach produces good similarity measures. In each case the similarity computation of our method is much better than traditional similarity approach showing the superiority.

Conclusion and Future Scope

The Semantic web which provides several instruments for improving search strategies and retrieving relevant web pages. The semantic similarity between the semantic web documents further improves the searching of relevant web pages. Also many similarity computation algorithms have been proposed to fully utilize the semantic annotations done and ontology-based concepts and relations.

The ontology based novel approach presented in the paper takes the ontology, and web page content into consideration to compute the similarity between the documents to the true value to improve the intended-search. Our future efforts would be to design more meaningful and exhaustive semantic web pages, so that the semantic search engine can evaluate more precisely relevance and also the similarity between the web page and retrieve them on taking any ontology already created or defining a new ontology by our approach. We will also try to make our approach scalable for the semantic web.

References

Berners-Lee T., Hendler J., and O. Lassila (2001), The Semantic Web, *Scientific Am*.

Brin S. and L. Page (1998), 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.

Chim H., and Xiaotie D. (2007), A New Suffix Tree Similarity Measure for Document Clustering, International WWW conference *ACM transactions*.

Ding L., Kolari P., Ding Z., and S. Avancha (2007), Using Ontologies in the Semantic Web: A Survey, Ontologies, *Integrated series of information systems*, vol 14, pp. 79-113, Springer.

Fernando S., and Mark S., A semantic similarity approach to paraphrase detection, Department of computer science, university of Sheffield, S1, 4DP, UK

Hajian B., and Tony W. (July 2011), 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.

Jan K. (Sept 2009), Systems for Discovering similar documents, Ph.D. Thesis proposal, Faculty of informatics Masaryk university.

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

Li Y., Bandar Z., McLean D., and James S.(2012), A method for measuring sentence similarity and its application to conversational agent, Proc. Of 17th *Int’l. Conf. FLAIRS*, Florida, USA, AAAI Press.

Nagwani N., and Shrish V (2011)., A frequent term and semantic similarity based single document text summarization algorithm *International Journal of Computer Applications*(0975-8887), vol-17, No. 2.

Oleshchuk V., and Asle P. (2003), Ontology Based Semantic Similarity Comparison of Documents, *Proc. of IEEE 14th workshop on database and expert systems* applications.

Page L., S. Brin, R. Motwani, and T. Winograd (1998), The Page Rank Citation Ranking: Bringing Order to the Web, *Stanford Digital Library Technologies* Project.

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

Protiti M. (2007), Semantic web: The future of WWW, Proc. Of 5th Int’l Conf. CALIBER, *Punjab University,* Chandigarh, 08-10.

Thiagarajan R., Manjunath G., and Markus S. (2008), Computing semantic similarity using ontologies ISWC 08, the I*nternational Semantic Web Conference* (ISWC), Karlsruhe, Germany.

Tho Q., Hui S., and Tru C. (2006), Automatic Fuzzy Ontology Generation for Semantic Web, *IEEE Transactions on Knowledge and Data Engineering,* Vol. 18., No. 6.

Varelas G., Voutsakis E., and Paraskevi R., Semantic similarity methods in Wordnet and their applications to information retrieval