Semantic Similarity between Documents Using Tree View Ontology

Anju Kundu[1], Mamta Kathuria[2]
[1, 2]Department of Computer Engineering, YMCA, Faridabad, Haryana, India

***Abstract*— *In Today's environment most of the web search engine are based on semantic check which provide help to find out meaningful documents from the web. They generally rely on keyword matching while returning document in response to user query. There are many documents that are related to each other in a semantic way. These documents are semantically same. Then keyword matching does not provide exact results in order to return these semantically similar documents in response to user query. This paper proposes a new method to find semantically similarity documents based on tree view ontology. The relation is enhanced by using relevant terms and calculates the fractions of these relations based on tree view. In this paper we provide performance results among different methods which shows the better affiance of proposed scheme.*

***Keywords*— *Document Similarity, Keyword matching, Semantic similarity, Tree View.*

I. INTRODUCTION

With the huge amount of information presented on the web, it has been difficult to access relevant information by the users of web and maintain the information by any machine. This is because web content is presented primarily in natural language, and targeted to human reader. However, some search engines such as Google, Yahoo etc. are being used by user in order to access the desired result. This type of desired information is based on similarity checking. "Similar" word is referring to a resemblance in appearance, character, or quantity, without being identical. To check the similarity among documents is known as document similarity. We can verify this similarity if they give information about identical subject or define similar meaning in the form of synonyms. But sometime traditional methods do not show similarity among documents. If they do not contain exact keyword or some form of semantic meaning. But the method which can automatically compute semantic similarity among documents is more helpful as compare to keyword matching method. Exactly measuring semantic similarity between text documents presents a major challenge due to the difficulty and ambiguity of natural language semantics. Generally documents are said to be similar if they are predicted to convey the same idea or subject and these synonyms are check from word to word then sentence to sentence and at last to document to document. But in case of no one of synonyms present between both documents, What we should do. Now there is a requirement of such method which shows some similarity if there is a real relation exists between them**.** The" relationship" word shows if both documents provide the information about same subject or subject are related in some manner. **Ontology** term defines "**The relationship exists between domains and it collect and organize terms of references**"[1]. Using ontologies we can increase the route by addition of the concepts that are not present in documents. This enhancement may make the documents similar which are not already similar before this increment. Synonyms can be used for parallel expansion and vertical expansion for subclass. This paper organizes the ontology in tree view format for calculating similarity. This method will provide semantic similarity using ontological tree view which will create through some enhancement in footprint.

1.2 Related Work

When a number of methods combine, these provide better results than single method but some of them perform well in specific application [1]. There is a problem with document similarity computation that simple methods are not so much reliable, the statistics methods based on additional corpus data, some methods based on surface similarity rather than on semantic similarity. The following table show some measures of related methods:

TABLE 1. Method description.

Measure Classes	**Class description**	**Main measures**	**Method description**
Binary similarity models	word-based, keywords based and n-gram measure to determine similarity	Tversky's Contrast Model[6]	Measures similarity as the ratio of common to common and distinctive features.
		Common Features Model[7]	Assumes simply that similarity common features
		the Distinctive Features based contrast Model[8]	Assumes that two motivations become more unlike to the extent that one motivation has a feature that the other does not
Count similarity models	Similarity models mainly based on the corpus representations using counts[9]	the Correlation model	Correlation measure
		the Jaccard model	Jaccard measure
		the Cosine model	Cosine-vector
		the Overlap model	the Overlap model
LSA similarity models	Latent Semantic Analysis models	the local weighting function[10]	Measures the importance of a word within a document
		The global weighting function [11]	Measures the importance of a word across the entire corpus of documents, normalized each word using the local weighting function; an inverse document frequency measure, an entropy measure.

There are four major measure classes which are used for identification of document similarity such as binary similarity, LSA similarity models, count similarity and ontology based similarity model. The existing ontology based methods are usually based on direct mapping from text to concept, so their similarity calculation is also based on literal similarity. Many existing ontology-based approach [2-5] calculate the similarity between concepts by using different aspects. One idea is to get similarity between ontology nodes by checking the intersection of both ontology graph nodes [12].

1.2.1 Binary similarity model

This model based includes word-based, keyword based and n-gram measure for checking similarity. This model is easy but not so much consistent because it based on exterior similarity not on semantic similarity. It is not so much appropriate on that time when natural language' people convey their similar opinion through different words

1.2.2 Count similarity model

This model based on a large text corpus for arithmetical computation but it is hard to attain such large text corpus in definite application [13]. It is same in case of LSA similarity model. It involve some similarity measure such as Correlation measure, Jaccard measure etc.

1.2.3 LSA similarity model

Latent Semantic Analysis model is based on local and global weighting function. These weighting function are used to generate weighted corpus representation and it is subjected to singular value decomposition.

A Relation Based Page Rank algorithm is proposed by Fabrizio L. et al. for Semantic Web search Engine. The author proposed a ranking strategy which concerns with the relation of keywords. The algorithm depends on data provided by queries. The page relevance is considered by using probability whether the page actually contains relation whose existence was supposed by user at time of query written.

Vladimir O. et al. provide the concept on Ontology Based Semantic Similarity Comparison of Documents. In his paper he considered ontologies in the form of knowledge structures that specify their properties and relations among them for knowledge withdrawal from the documents. They represented ontologies using a graph-based model that shows semantic relationship among documents. Instead of doing raw document comparison there is comparison after enhancement of document.

II. Proposed Method

To overcome the shortcoming of some similarity methods, we use ontology based method. An ontology is referred to a vocabulary that describes a domain of interest and specify the meaning of terms used in that specific vocabulary [14].We use ontology tree view along with similarity checking at each root node along path. So we provide the name of this method is semantic similarity using ontology tree view.

The results are based on following calculations:

□ At The start of checking there is a keyword matching at root.

□ If keyword matching method fails then we apply semantic checking upon the enhanced tree.

□ If it will not work well then we apply checking on levels. If result is estimated to 0 then we can find that there is no connection present between both documents.

2.1 Ontology Tree View

Our main aim is to check how much similarity occurs between two or more documents. We select a **stored** document and other document which is **input** document for comparisons. We will create ontology tree view for both documents for comparison made at level to level and enhanced both tree view through their domain word's reference. Some steps which are involved for calculating semantic similarity based on ontological tree view. These steps are:

- Provide both documents at first level.
- Apply NLP parser and stopword removal procedure.
- Comparison starts from root (Firstly apply keyword matching method and if there will be failure of keyword match then check with semantic similarity).

If root node of stored doc will not match with root node of input Doc then root of stored Doc will be checked by nodes at sublevel in input Doc and we will repeat all iteration until all nodes will finish or no matching will occur as result. In case of no matching found we can say that there is no similarity present.

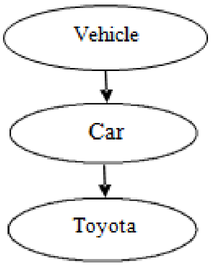


Fig. 1. Tree of document 1.

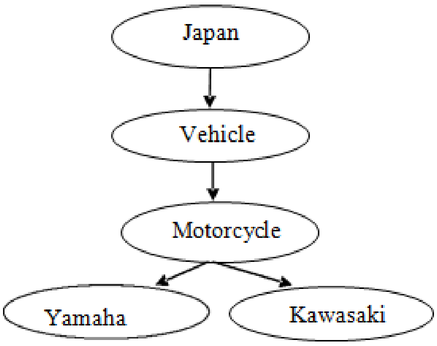


Fig. 2. Tree of document 2.

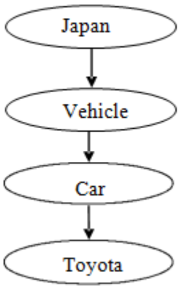


Fig. 3. Tree of document 1 after expansion.

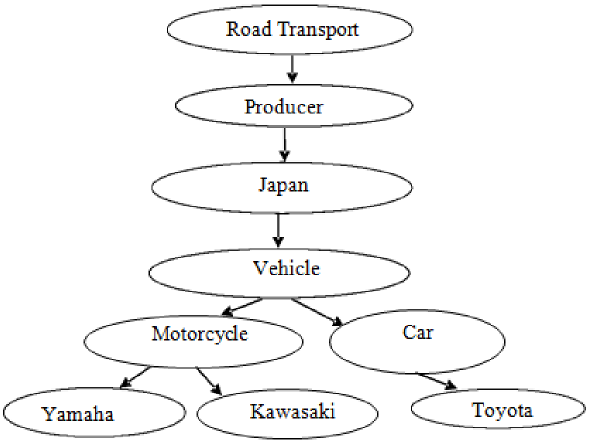


Fig .4. Ontology after expansion.

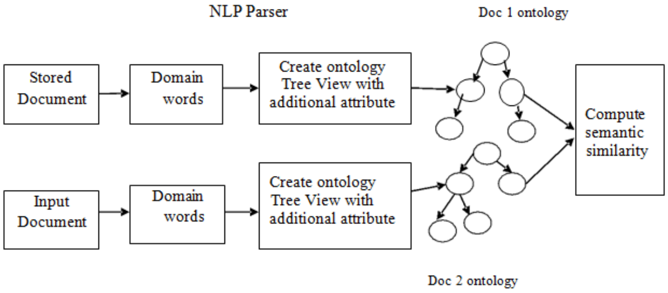

NLP Parser
Doc 1 ontology
Stored Document
Domain words
Create ontology Tree View with additional attribute
Input Document
Domain words
Create ontology Tree View with additional attribute
Compute semantic similarity
Doc 2 ontology

Fig. 5. Steps of semantic similarity using ontology tree view.

The documents are parsed by using NLP Parser with the help of Word Net then the trees of both documents are created and at last there is an expansion of data by using ontology for checking similarity. After expansion of tree in ontology form, nodes represent the object and edge represent relation.

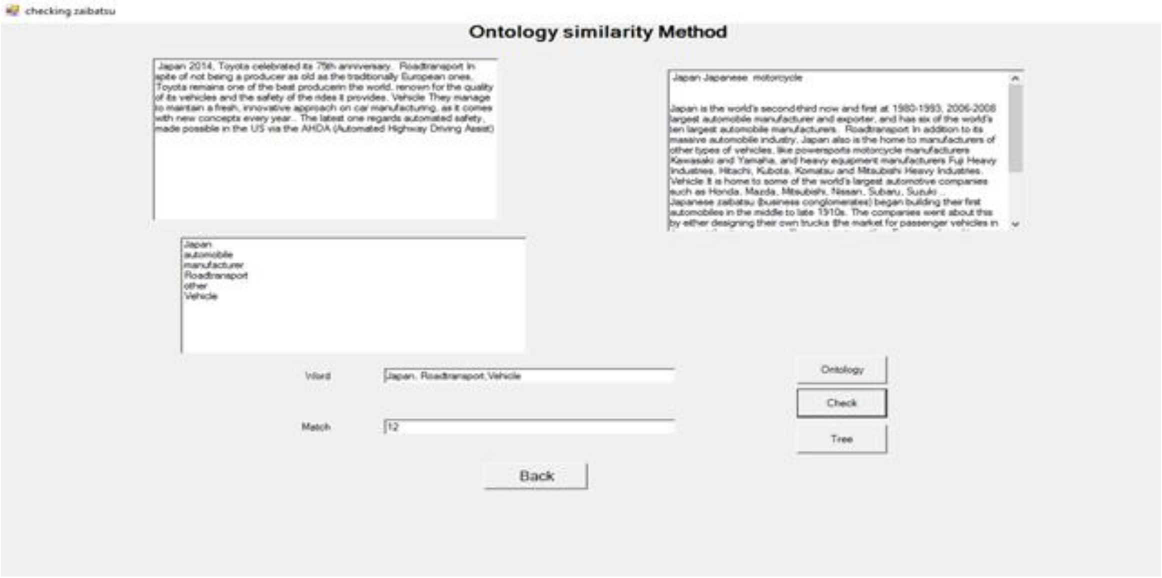
III. EVALUATION

To evaluate the performance of proposed method on semantic similarity, some experiments are performed. The result sets are based on percentage getting from different methods. These methods used vector space model for calculating similarity between documents. The high percentage show the efficient results which is given by proposed method.

3.1 Experiments Results

In original system, we have two documents depends on the selection of user for mapping the semantic similarity based on ontology tree view. We now examined the results produced by different methods. Firstly the keyword matching method gives results when there is a comparison between both documents. The result is 4 percent in this case which is specified by implemented system. The second method is semantic similarity which is based on semantic words. The checking of semantic words is done through wordnet which contains a lot of synonyms of each word present in the document. This method provides better results as compare to keyword matching method. This method gives 6 percent result in output which is better than keyword matching.

Now the proposed method is semantic similarity based on ontology tree view provides better results as compare to both methods shows its efficiency. It provides better results after using the additional attributes which are recognized as those terms which create relations between both documents by looking on both documents work as ontology. This method shows highest result which is 12 percent. These results can be shown in fig 6.



Japan 2014, Toyota celebrated its 75th anniversary. Roadtransport In spite of not being a producer as old as the traditionally European ones, Toyota remains one of the best producerin the world, renown for the quality of its vehicles and the safety of the rides it provides. Vehicle They manage to maintain a fresh, innovative approach on car manufacturing, as it comes with new concepts every year.. The latest one regards automated safety, made possible in the US via the AHDA (Automated Highway Driving Assist)

Japan Japanese motorcycle

Japan is the world's second third now and first at 1980-1993, 2006-2008 largest automobile manufacturer and exporter, and has six of the world's ten largest automobile manufacturers. Roadtransport In addition to its massive automobile industry, Japan also is the home to manufacturers of other types of vehicles, like powersports motorcycle manufacturers Kawasaki and Yamaha, and heavy equipment manufacturers Fuji Heavy Industries, Hitachi, Kubota, Komatsu and Mitsubishi Heavy Industries. Vehicle It is home to some of the world's largest automotive companies such as Honda, Mazda, Mitsubishi, Nissan, Subaru, Suzuki ..
Japanese zaibatsu (business conglomerates) began building their first automobiles in the middle to late 1910s. The companies went about this by either designing their own trucks (the market for passenger vehicles in

Japan
automobile
manufacturer
Roadtransport
other
Vehicle

Word: Japan, Roadtransport,Vehicle

Match: 12

Fig. 6. Semantic similarity using ontology.

The graph shows the performance of all methods which satisfy with the highest performance of our proposed method. At x axis stored document is present and on y axis other input document is present. The performance of all method is shown by different colors.

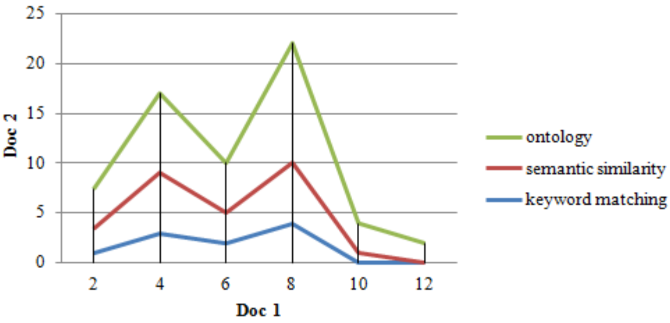


Fig. 7. Performance chart.

IV. Conclusion

This paper proposed a method which transforms the documents in a personal ontology. Here ontology shows that there is a relation presents between the documents in reality or not. The comparisons of all methods show that the proposed method provides the highest similarity results. This provides better results in case if there is a relationship exists otherwise no improved results than other methods.

V. Futurework

In this we will make tree view and compare it with root node with vector space method and find out some similarity between documents. But we will do it by using domain words find out through parse tree. Its analysis can be made easy if we will use some tool for tree view in future. So we will implement tree view ontology dynamically directly through domain keywords of stored and input file using some tool so that we can import directly tree view data from database in future.

Reference

[1] V. Oleshchuk and A. Pedersen, "Ontology based semantic similarity comparison of documents," *14th International Workshop on Database and Expert Systems Applications (DEXA'03)*, 2003.

[2] M. Batet, D. S´anchez, and A. Valls, "An ontology-based measure to compute semantic similarity in biomedicine," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 118–125, 2011.

[3] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871–882, 2003.

[4] E. G. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, "X-similarity: computing semantic similarity between concepts from different ontologies," *Journal of Digital Information Management (JDIM)*, vol. 4, no. 4, pp. 233–237, 2006.

[5] G. Pirr´o, "A semantic similarity metric combining features and intrinsic information content," *Data & Knowledge Engineering*, vol. 68, no. 11, pp. 1289–1308, 2009.

[6] A. Tversky, "Features of similarity," *Psychological Review*, vol. 84, issue 4, pp. 327–352, 1977.

[7] R. N. Shepard, and P. Arabie, "Additive clustering representations of similarities as combinations of discrete overlapping properties," *Psychological Review*, vol. 86, issue 2, pp. 87–123, 1979.

[8] D. L. T. Rohde, "Methods for binary multidimensional scaling," *Neural Computation*, vol. 14, issue 5, pp. 1195– 1232, 2002.

[9] M. E. Rorvig, "Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets," *Journal of the American Society for Information Science*, vol. 50, issue 8, pp. 639–651, 1999.

[10] M. D. Lee and D. J. Navarro, "An empirical evaluation of models of text document similarity," *The annual conference of Cognitive Science Society*, pp. 1254-1259, 2002.

[11] B. M. Pincombe, "Comparison of human and latent semantic analysis (LSA) judgments of pairwise document similarities for a news corpus," *Defence Science and Technology Organisation Research Report DSTO–RR– 0278*, 2004.

[12] C. Xie, M. W. Chekol, B. Spahiu‡, and H. Cai, "Leveraging Structural Information in Ontology Matching," *IEEE 30th International Conference on Advanced Information Networking and Applications (AINA)*, 2016.

[13] R. Shams, A. Elsayed, and Q. Mah-Zereen Akter, "A Corpus-based evaluation of a domain-specific text to knowledge mapping prototype," *Journal of Computers*, vol. 5, issue 1, pp. 69-80, 2010.

[14] P. Shvaiko and J. Euzenat, "Ontology matching: State of the art and future challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, issue 1, pp. 158-176, 2013.