Relation based Measuring of Semantic Similarity for Web Documents

Poonam Chahal
Research Scholar
YMCAUST, Faridabad

Manjeet Singh
Associate Professor
YMCAUST, Faridabad

Suresh Kumar
Professor, FET,
MRIU, Faridabad

ABSTRACT
The World Wide Web (WWW) is the information resource centre in which information exists in the structure of web pages which are interlinked with each other. From the huge amount of information present on WWW it has been found difficult to extract the relevant information for the query given by the user. The reason for this is that the information exists on web is in natural language. The layered architecture semantic web is given by Tim Berner Lee to overcome the issues of information retrieval. In recent times, numerous semantic web search engines have been developed like Ontolook, Swoogle, etc which assist in searching significant documents presented on semantic web. Several attempts have been made in ruling out the similarity of semantic web pages but then also the results of these semantic similarity techniques between web documents is neither appropriate nor upto the user's prospects. This paper proposes an approach for finding the semantic similarity between the web documents along with the consideration of the concepts as well as the relationships that will exists between the concepts also. In our approach the documents are being processed by extracting concepts and relationships between the existing concepts from the documents using the base ontology and the dictionary having the words along with the synonyms. Finally, the set of any two documents are compared to find their semantic similarity by taking the relationships that exists in the documents. We discover all relevant relationships between the words which provide the core information of the document and then the similarity of these relationships is computed on each web page to find out their significance.

Keywords
WWW, Information, Retrieval, Ranking, Semantic, Similarity.

1. INTRODUCTION
This paper introduces a method for measuring similarity between web documents using ontology. Ontology is the specification of conceptualization in context of knowledge [10]. Similarity can be categorized based on attributes and relations. The attributional similarity corresponds to compare the attributes and in contrast to this the relational similarity compares the relations. The high

Degree of attributional similarity represents the words are synonyms and the high degree of relational similarity represents the words analogous.

WWW (World Wide Web) is considered as the information resource center [1]. So, for retrieving relevant information from www it is necessary to find semantic similarity which helps in providing the relations and instances of relations that exists between two words/sentences/documents. The method proposed in this paper helps to retrieve the semantic relationships so that better result set can be produced for a query provided by the user of search engine.

2. RELATED WORK
As there is large amount of information present in WWW, it has been a very crucial process to extract relevant information [11]. The crawling, indexing and ranking of documents by a search engine can be mainly done by finding the semantic similarity between the web pages. Many researchers have tried to explore the idea of finding the semantic similarity but still the result-set provided by existing similarity detection methods are not up to the user expectations.

In our paper we are providing a different approach of finding the semantic similarity between the documents which not only considers the concepts/terms that exists in the documents but also the relationship between the concepts.

In contrast to the semantic matching there exists the lexical matching approach in which only words are considered [2]. The vector space model (VSM) is used in lexical matching. In lexical matching for each document the vector space model is created consisting of the words which are extracted from the documents and then the similarity is computed using Cosine similarity, Jaccard similarity, Dice similarity, etc [12]. Many researchers have provided different methods for finding the semantic similarity but to integrate the semantic feature in the search engine needs its growth in the form of layered architecture to hold semantic web focuses on taking into consideration the concepts and relationships between the concepts that subsist in the document [4].

Bollegala D. et. al. [3] proposed measuring the similarity between implicit semantic relations from the web. In the paper the author provides the method for semantic similarity by including three components. The components that the authors considered are representing the various semantic relations that subsist between a pair of words using automated extracted lexical patterns, clustering the extracted patterns to identify different pattern, and measuring the semantic similarity using a metric learning approach.

Lee Chee M. et. al. [7] gives the grammar based semantic similarity algorithm for sentences. The author proposed a algorithm for sentence similarity which takes into consideration the benefit of corpus-based ontology and grammatical set of laws to overcome the problems addressed. The authors also give experiments on two famous benchmarks which demonstrate that the algorithm proposed by them has a major performance improvement in sentences/short-texts with random syntax and structure.

Gan M. et. al. [8] discussed on calculation of semantic similarity based on ontology. The authors classified the existing methods into five categories: semantic distance based methods,

information content, properties of terms based methods, ontology hierarchy, and hybrid methods. Then the authors summarized characteristics of each category along with advantages and disadvantages of these methods.

Liu H. et. al. [5] proposed assessing text similarity using ontology. The authors presented a novel approach for assessing similarity by finding linkage between ontology terms and sentence. These are used for the generation of semantic indexes for sentences or document and this is applied by the authors with a searching algorithm to compute semantic similarity.

In all the above contributions by different researchers the main center of attention is on establishing the semantics either by taking concepts or relationship that exists between any two concepts in the ontology. It is essential to evaluate the complete semantic similarity among the documents to find the factual value of similarity between the documents [6].

3. PROPOSED SEMANTIC SIMILARITY MODEL FOR WEB DOCUMENTS

In our proposed model the semantic similarity between the documents is found by first extracting the words from the documents and then the concepts represented by them using a domain specific dictionary. Also a weighted ontology is built related to a specific domain in which nodes represent the most common words/concept and the edges represent the relationships that can exists between the nodes. The relationships that exist between two concepts are assigned weights in the ontology depending on the class and subclass relationships between the concepts. The domain specific dictionary in which words/term/concept related to a domain is present along with the synonyms that can exist for them. This dictionary is prepared using WordNet and the base ontology is constructed using Protégé.

Given two pair of concepts (C1, C2), (C3, C4) presented in document D1, D2, the method of measuring semantic relationships that exists between the concepts can be done in two stage.

First, we have to retrieve lexical pattern that exists in each pair of concepts present in D1, D2 using search engine. For lexical pattern retrieval we need to extract various contexts in which two concepts in a concept-pair co-occur using the base ontology O. The web search engine is used to extract the text snippet for the context of two concepts. Snippets are the summaries provided by the search engine like Google, Yahoo etc. in the search results. To retrieve the snippets for the concept pair A, B, we can provide the seven types of queries i.e. A*B, B*A, A**B, B**A, A***B, B***A, and AB. Here the * corresponds to the returned snippet.

Second, the relationships between the concepts are extracted using the base ontology O. The relationships that exist in the ontology are assigned weights according to their type and class-instance relationships between the concepts. These extracted relationships are used to compute the Relation space model (RSM) for each document. The RSM is constructed as the vector space model used for lexical matching between documents. The RSM of a document is having the concept pairs and the relationships along with the frequency of the relationship.

Finally the RSM and the lexical patterns are used to compute the semantic similarity between the concepts present in the document.

The semantic similarity is computed by considering a document and analyzing the document syntactically and semantically using LGP Parser. This help in retrieving the words present in a document. These words are replaced by concepts representing them using the domain specific dictionary. Then, the relationship between the concepts is established along with the frequency of that relationship for the document using the RSM constructed for a document. These relationships are sorted by using the frequency of a relationship that is existing in a document. The more the frequency of the relationship more meaningful it is for the document. Similarly other document RSM can be created having the information of all relationships between concepts pairs present in a document along with the frequency. The lexical patterns retrieved between the concepts pairs are used to compute the lexical similarity using cosine similarity.

Finally, the lexical matching score and the RSM computations are used to compute semantic similarity between the documents.

Semantic similarity score formula:

$$SSc = (RSMfij + LMij)/2 \quad (1)$$

Where RSM_{fij} is the relationship cumulative frequency weighted score for document i and j.

LM_{ij} is the score obtained from the lexical matching of document i and document j.

The model for our approach is shown in figure1. The main component of the model are ontology processor, document processor, semantic score computation module. The ontology processor is used to extract the concepts and their relationships from the set of documents using concept analyzer and relation analyzer. The document processor is extracting words from the set of documents using LGP Parser which can be represented by concepts present in domain specific dictionary integrated with the help of WordNet. The computations retrieved from the RSM and lexical are given to semantic score computation module to obtain the semantic score of a document.

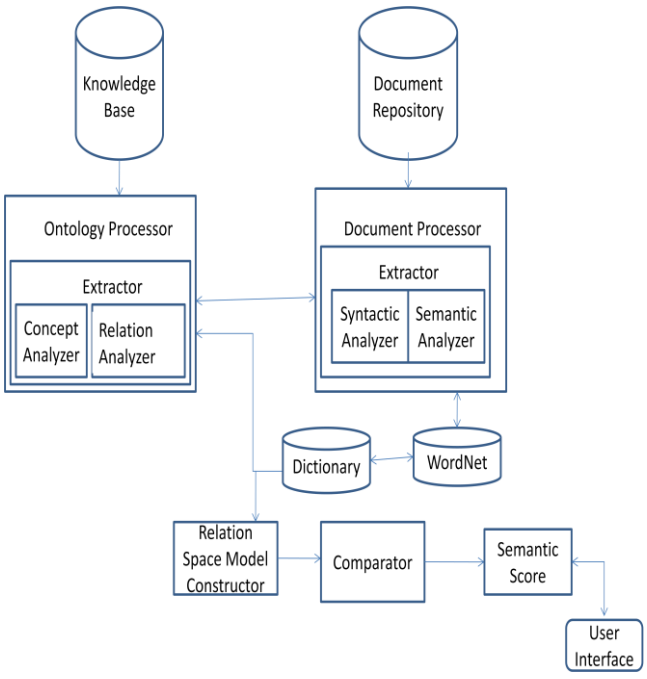

Fig 1: Structural design of proposed Semantic Similarity Model.

The algorithm for our approach shown in figure 1 is as follows:

1. Construct a Domain specific weighted Ontology O.
2. Construct a Domain Specific Dictionary D having keywords/terms along with the synonyms.
3. For each document in document set do
   i. For each sentence in the documents Di extract keyword/term/concept ti.
   ii. Search the term ti in D to find the most popular synonym ci that is also present in the base ontology O.
   iii. Replace ti with ci.
   iv. Extract the relations ri that exists in the document with respect to O.
4. For two documents Di, Dj do:
   i. Construct the document RSM (relation space model) having relationships that exists in document along with frequency of the relationship ri.
   ii. The RSM constructed is sorted according to the frequency of relationships present in the space model and normalized.
5. Compare the relation space model of both the documents by taking the cumulative weights of the matched relations of both the relation space model of the documents.
6. Finally compute the semantic score between two documents using equation 1.

In this section the working of our approach with the help of examples is given. In our analysis we are taking documents D1 and D2 related to domain computers and the document ontology for D1 and D2 is shown in figure 2.

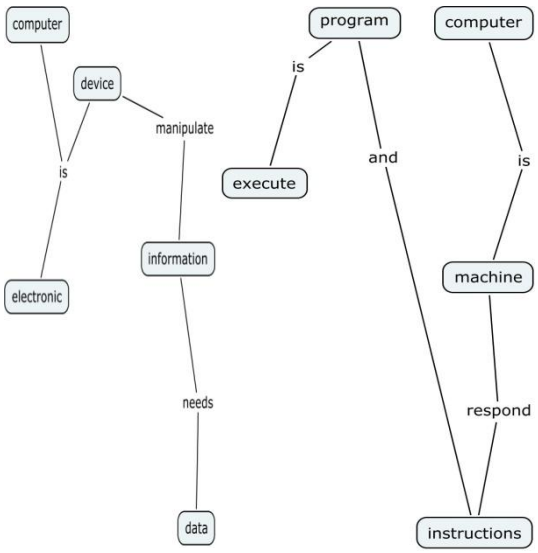

Fig 2: Document ontology for D1 and D2 using RSM.

D1: A computer is an electronic device that manipulates information or data.

D2: A computer is a machine that responds to specific set of instructions and executes program.

From these two documents extracted words are shown as nodes and extracted relationships are shown as edges. Now these words are replaced with concepts using domain specific dictionary D as shown in Table 1. These concepts are used to create the relation space model for the document using ontology shown in Table 2.

Table 1: Dictionary Based Weights

S No	Words	Synonyms as concepts
1	Computer	Machine, device, -,-,-,-,
2	Study	Survey,work,report,-,-,-,-,
3	Machine	Device, product, mechanism,--,-,-,
4	Science	Branch, discipline, field,-,-,-,-,
5	Intelligent	Ability, knowledge,-,-,-,-,
6	Design	Plan, blueprint,-,-,-,-,-,-,
7	Agent	Factor, broker,-,-,-,-
8	System	Scheme,-,-,-,

Table 2: Ontology Based Weights

S No	Concept-Relationship Between Words from Document Repository	Weights Assigned
1.	is(computer, machine)	1
2.	Has(system, agent)	.8
3.	Has(person,education)	.6
4.	Needs(machine,science)	.7
5.	Needs(program,instructions)	1
6.	Isa(machine, device)	1

The relation space model is then normalized and the cumulative frequency weights are considered for computing semantic similarity score using equation 1.

4. PERFORMANCE ANALYSIS

Performance of proposed method for computing semantic similarity between the web documents using ontology depends not only on keywords but also the associated relationships between the keywords/concepts/terms that are extracted from the document using ontology. We have computed the performance of the corpus having 50 documents and the dictionary having words with the synonyms are from these domain specific documents.

We have evaluated the performance of our proposed semantic similarity method with the similarity computed by the vector space model approach and Euclidean approach [3]. In our approach we are finding the semantic similarity between concepts present in the document using relations is giving improvised results.

Table 3: Results of similarity VSM, EUC and proposed semantic similarity approach.

S no.	Set of Documents	VSM	EUC	Proposed
1.	D_1,D_2	.5	.6	.7
2.	D3,D_1	.4	.49	.8
3.	D4,D_5	.49	.6	.6
4.	D2,D3	.55	.57	.71
5.	D3,D4	.32	.36	.36
6.	D1,D4	.44	.47	.55

Note: D1: http://en.wikipedia.org
D2: http://www.tutorialspoint.com
D3: https://class.stanford.edu
D4: http://www.webopedia.com
D5: http://www.techopedia.com

For deep analysis for the performance of our proposed approach with VSM [9], we further searched for the pages from Google search engine which are having similar content but using different words. The maximum number of PDF files having similar content was taken and the ontology based approach applied to the documents. We found that for utmost number of documents the proposed approach creates good similarity score. For each case the similarity calculation of our technique is much better than conventional similarity approach showing the pre-eminence.

5. CONCLUSION AND FUTURE SCOPE

The semantic comparison between the web documents additional improves the searching of relevant web pages. Many similarity calculation algorithms have been proposed to completely utilize the concepts and relationships that exists between the concepts.

The approach presented in our paper takes the ontology, and content presented in web page to calculate the similarity between the documents to get better result-et for a query given by the user. Our upcoming efforts would be to devise more significant and extensive semantic web pages, so that the semantic similarity between the web pages can be computed efficiently using ontology already formed or construing a new ontology.

6. 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] Danushka B., Matsuo Y., and Ishizuka M., " Measuring the Similarity Between Implicit Semantic Relations from the Web", International World Wide Web Conference Committee ACM, pg 651-660, 2009.

[4] Ding L., Kolari P., Ding Z., and S. Avancha, "Using Ontologies in the Semantic Web: A Survey", Ontologies, integrated series of information systems, vol 14, pp. 79-113, Springer, 2007.

[5] Hongzhe Liu and Pengfe Wang, "Assessing text semantic similarity using ontology", Journal of Software, vol 9, No. 2, Feb 2014.

[6] Lee W, Shah N., Sundlass K., and Musen M.,"Comparison of Ontology Based Semantic Similarity Measures", AMIA Annu Symp Proc., pg 384-388, 2008.

[7] Lee M., Jia C., and Tung H., "A Grammar Based Semantic SimilarityAlgorithm for Natural Language Sentences", Research Article in Scientific World Journal, Pg(s) 17, vol 2014.

[8] Mingxin G, Duo Xue and Jiang Rui," From Ontology to Semantic Similarity: Calculation of Ontology Based Semantic Similarity", Scientific World Journal, pg(s) 11, Vol 2013.

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

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

[11] Page L., S. Brin, R. Motwani, and T. Winograd, "The Page Rank Citation Ranking: Bringing Order to the Web", Stanford Digital Library Technologies Project, 1998.

[12] 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.