Ranking of Web Documents using Semantic Similarity

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

*Research Scholar, YMCAUST, Faridabad ** YMCAUST, Faridabad, ***FET, MRIU, Faridabad
Poonamnandal.fet@mriu.edu.in, mstomer2000@yahoo.com, enthusk@yahoo.com

***Abstract*- In recent years, semantic search for relevant documents on web has been an important topic of research. Many semantic web search engines have been developed like Ontolook, Swoogle, etc that helps in searching meaningful documents presented on semantic web. The concept of semantic similarity has been widely used in many fields like artificial intelligence, cognitive science, natural language processing, psychology. To relate entities/texts/documents having same meaning, semantic similarity approach is used based on matching of the keywords which are extracted from the documents using syntactic parsing. The simple lexical matching usually used by semantic search engine does not extract web documents to the user expectations. In this paper we have proposed a ranking scheme for the semantic web documents by finding the semantic similarity between the documents and the query which is specified by the user. The novel approach proposed in this paper not only relies on the syntactic structure of the document but also considers the semantic structure of the document and the query. The approach used here includes the lexical as well as the conceptual matching. The combined use of conceptual, linguistic and ontology based matching has significantly improved the performance of the proposed ranking scheme. 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 with respect to the query provided by the user. We have found that this semantic similarity based ranking scheme gives much better results than those by the prevailing methods.**

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

I. INTRODUCTION

Information retrieval has been the topic of research for last few decades. 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 [13]. With the growth of internet users looking to the information presented on the web are continuously facing the difficulty to find or access relevant information or maintain the information on any machine. The reason for the difficulty of extracting the information is because web content is presented primarily in natural language, and targeted to human reader. However, the information retrieval tools, such as Google, Yahoo etc. are being used by human reader in order to access the desired information. But, the result-set produced by the search engines with respect to the query of the user are not upto the user expectations as it includes many irrelevant web pages in the list of the ranked documents. This is because the technique used by the majority of the traditional search engines is based on the lexical matching and link analysis [12], which have inherent defects.

In recent years, the researches on semantic similarity have explored the idea that it is not sufficient to search the text presented on the web only by keyword matching. There are documents presented on the web that means the same thing but use different words which are having same meaning. To overcome this limitation of lexical matching it is necessary to have the approach of finding the semantics of the document by taking not only the keywords extracted by the syntactic parsing but also the relationship that exists between the keywords.

The semantic web visualized by Tim Berners-Lee [1] 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. Basically, the ranking of the documents retrieved is done by the search engine which include list of relevant documents rather than irrelevant can be efficiently done by finding the semantic similarity between the document and the query specified by the user to retrieve the information from the web. Some attempts have been made in ranking the documents by finding semantic similarity between the documents and the query given by the user to the search engine but still the results provided by similarity detection techniques are not up to the user's expectations.

Although there are various techniques based on Lexical analysis, Natural Language Processing, Ontology based matching, Semantic analysis, etc. for computing the similarity of the documents and the query given by the user. In this paper

978-1-4673-5986-3/13/$31.00 ©2013 IEEE

we have tried to explore a novel ranking model which provides the result-set according to the user query by considering their relevance by keeping the user view in mind and also the semantics of the document and the user query. In section II we have discussed the related work, further in section III the detailed proposed ranking model is given. The performance analysis of the proposed ranking model is given in section IV, and finally in section V we have given the conclusion and future scope.

II. Related Work

In fact, the information retrieval process by a search engine is very crucial. There are many ranking models [7] that have been proposed by the various researchers like Boolean model [6], statistical model [9], Hyperlink based model [3], Conceptual model [16] and many more [4, 8] which has been widely used. Some of them use the natural language processing techniques such as language model and relaxation algorithm. The use of natural language techniques in these models helps to consider syntactic, semantic structure and morphological form of terms. Some other document ranking models such as Neural Networks, Fuzzy Sets, Relevance Feedback Models could be used for efficiently increasing the performance of ranking models.

Zhanun. et. al. [10] have given Ontology-based design information extraction and retrieval. In this work the authors used the natural language processing and domain-specific ontology to automatically construct a structured and semantic based representation from the unstructured documents. The concepts and relationship between the concepts are recognized with the help of linguistic patterns. Further, the concepts and relationship are represented in the form of conceptual graph. Finally, the integration of these conceptual graph builds the domain specific ontology which can be compared to any other ontology for finding the semantic similarity between the two documents.

Mehrnoush Shamsfard et. al. [14] have given a method of ORank: An Ontology Based System for Ranking Documents. In their paper the authors proposed the new method of ranking by determining semantic similarity based on structure-knowledge extracted from ontology. The approach given by the authors exploits natural language processing techniques for extracting phrases and stemming words. Then the authors used the ontology based conceptual method to incorporate the semantics by annotating the documents and also expand the query. The spread activation algorithm is used and improved for the expansion of the query so that it can be done in various aspects. Finally, the annotated documents and the query which is expanded can be used for processing to compute the relevance degree by exploiting the statistical methods.

Danushka Bollegala et. al. [2] proposed a relational model of semantic similarity between words using automatically extracted lexical pattern clusters from the web. The authors in this paper first represented the semantic relation that the two given words holds by using automatically extracted lexical pattern clusters. Then the authors used Mahalanobis distance measure to compute the semantic similarity between the words. Their proposed relational model differs from the feature model of the similarity used by others as it is defined over the set of semantic relations that exists between the words instead of the set of feature for each word.

Jiwei Zhong et. al. [18] proposed conceptual graph matching technique for semantic search. In their paper the authors have given the approach of semantic search by the help of matching the conceptual graph. In this approach the description of web pages is extracted by the automating technique like wrapper induction. After that each description is converted into a conceptual graph using ALPHA system. The conceptual graphs generated are then to be stored in the resource CG repository. The CG matching handler module is given the input consists of one query graph and one candidate graph fetched from the resource CG repository and the ranking of the candidates is returned to the user interface in the form of the output.

In all these contributions given by various researchers the focus is on introducing the semantics either by taking ontology [5] or relationship that exists between the concepts. To make ranking of the documents by the search engine having only the relevant pages in the result-set it is necessary to compare the complete semantic similarity between the documents and the query given by the user. We have attempted this and have achieved very good results.

III. Proposed Ranking Model for Semantic Web Documents

To retrieve relevant documents from the web efficiently the ranking done by the search engine should be able to find the semantic similarity between the query and documents by keeping the user view in mind, i.e. by considering the query and extending the query using ontology [11]. Thus we can construct ontology for any document [15, 17].

In this paper we have developed a novel approach for ranking of the documents by finding the semantic similarity between the semantic web documents and the query given by the user for efficient information retrieval. The overall system architecture is given in Figure 1. The main components of the architecture of the system are ontology processor, ranker module, document processor. In the approach given here we first determine the keywords of the document using syntactic analysis and making the vector space model of the documents. For this a domain-specific dictionary is prepared having the words, their synonyms along with their meaning which are related to the domain. The words present in the dictionary are assigned weights based on their relevance/relation with the domain using fuzzy set approach. The query which can be given related to the domain have been framed using the standards of the Wordnet. We have thus developed a database of words, their weightage, their synonyms, ontology etc. We have represented this database as a vector space model for the processing purpose.

The mapping of each words present in the vector space models stored in the document repository is done with the domain specific dictionary weighted terms called semantic dictionary database. This is done sentence wise. Each sentence will contain a relevance value and then the integration of all the sentences is done by using statistical approach. This integrated relevance value obtained above will form the relevance of the individual paragraphs. Finally, the relevance value of the document is obtained by integrating the relevance score of all the paragraphs present in the document again by using statistical model.

The relevance score of the document obtained above is for the query specified for the domain as the mapping of vector space model is done with the help of fuzzy weighted terms in the domain specific dictionary stored in dictionary repository.

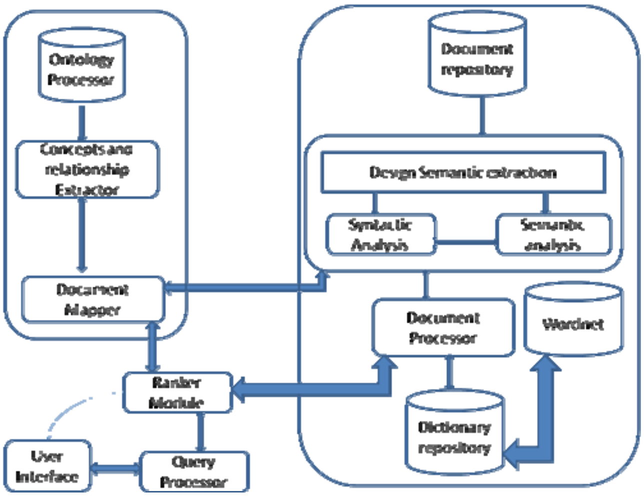

Figure 1: Architecture and System Flow Diagram of proposed Ranking Model.

The document similarity with respect to the user query is also found by extracting the concepts and the relationship that exists between the concepts present in a document and the mapping of the same is then done by the ontology processor. The ontology processor generates all the weighted relationship that are present in the ontology and the whole database is mapped with the document vector space model. The score obtained by this mapping is with the help of statistical model. Finally the maximum value gives the overall relevance of the document with respect to the query. The algorithm for our method is as follows:

Algorithm

Step1: Create a Text-List (by links).

Step 2: Take query as a text: a String.

Step 3: For each Text in Text-List do:

(a) Construct Text-Vector-Space.
(b) Construct Domain-Dictionary of words.
(c) Using Statistical-Model() and Domain-Dictionary, Calculate relevance-value of Text with respect to Query.
(d) Construct Domain-Ontology of the Text.
(e) Calculate Domain-Similarity of Text value with Domain-Ontology.
(f) Determine the maximum of Domain-Similarity value and relevance-value and call it Relevance-Score.

Step 4: Goto step 3 until no text is left in Text-List or no more texts are to be considered.

Step 5: Arrange the text (links) according to decreasing order of relevance-score and assign them ranks.

Step 6: Display the texts according to their ranks.

Text-Vector-Space: Consists of text words their weightage.
Domain-Dictionary: Consists of text-words (nouns, pronouns, synonyms and their weightage).
Domain-Ontology: A graph containing concepts as nodes and relations as edges.
Domain-Similarity is calculated for the Text with respect to Domain-Ontology and Domain-Dictionary.
Statistical-Model is used to calculate the relevance score of text with respect to domain-Dictionary.

Further, we will explain the detailed working of our model with the help of examples. We have taken four documents related to the education domain. The domain specific dictionary database is made with the help of WordNet and then fuzzy weights are assigned according to their relevance to education. A part of the database is shown in Table 1. Also the database of relationships between the concepts related to the domain is constructed with the help of ontology in which the nodes represents the concepts and relationships between the concepts are represented by edges between the nodes. We have not shown the graph here but the extracted database from the graph following ontology processor is given in Table 2.

Table 1: Dictionary Based Weights

S No	Domain (Education)	Weight Assigned	Synonyms
1	Process	.8	-,-,-
2	Study	1	-,-,-
3	Learning	1	-,-,-
4	Experience	.8	-,-,-
5	Social	.9	-,-,-
6	Official	.8	-,-,-
7	Dynamic	.8	-,-,-
8	Starting point	.5	-,-,-
9	Used every where	1	-,-,-
10	University	.6	-,-,-

The documents taken for analysis of our ranking model with the help of example is shown in Table 3. Now, in the working of the example the query used is “what is education”. The user query used is related to the domain education whose dictionary is stored in the dictionary repository.

Table 2: Ontology Based Weights

S No	Concept-Relationship Between Objects Represented in FOL	Weights Assigned
1.	Of(education,man)	.8
2.	Relatedto(education,study)	.7
3.	Has(person,education)	.6
4.	At(education,college)	1
5.	At(education,school)	1
6.	At(education,university)	1
7.	Isa(education,process)	1
8.	Has(life,learning)	.9
9.	Through(learning,experience)	.9

Now, when the document is parsed using natural language processing techniques then its relevance value is calculated with respect to the query point of view by considering the dictionary and the ontology. Finally, the max value is assigned to the documents to give their final relevance value as shown in table 3. According to the relevance value obtained, the documents are ranked by keeping the user view in mind, so that the user gets the relevant pages with respect to their respective query.

The same set of documents are ranked with the Google and the ranked value obtained with Google Page Rank search are .62, 1.24, 1.11, 1.13 for documents D1, D2, D3 and D4 respectively. We have also taken the human ratings for the same set of documents and given the weightage accordingly to all the documents. Finally the variance is calculated for both the search engine i.e Google and the proposed method of ranking with respect to actual rank score given by humans. We find that the variance calculated by the proposed ranking method is minimum as compared to the variance calculated by the Google ranks showing its superiority.

Table 3: Comparative Relevance of Documents to User Query

SNO	Document No.	Relevance value to domain specific dictionary (Dv)	Relevance value to ontology (Ov)	Final Relevance value=Max(Dv, Ov)
1.	Document1	.85	.94	.94
2.	Document2	.74	.88	.88
3.	Document3	.45	.75	.75
4.	Document4	.6	.65	.65

D1: Education is a lifelong process. A person learns through his experience. It goes on forever from his birth to death without any break or barrier.
D2: Education of man does not begin at school but begins at birth. It ends not when he graduates from university but ends at his death. Hence, Education is a lifelong process.
D3: Education is not only academics but social also. It is important in one’s person life.
D4: In a person life everyone needs to be educated and social. Everyone learns through experiences gained in one’s life.

We have taken two other set of documents represented by set (D21, D22, D23, D24, D25, D26) and set (D32, D33, D34, D31) respectively for further analysis. The overall result analysis is given in Table 4. The contents of these documents are as follows:

D21: Education in its broadest sense is the means through which the aim and habit of a group of people lives on from generation to generation.
D22: Education means the process of becoming an educated person.
D23: Education means to know the knowledge.
D24: Education teaches lesson of humanity. It is very necessary for humans.

D25: Education is the act or process of imparting or acquiring particular knowledge, as for a profession.
D26: Education psychology involves the study of how people learn.
D32: Education is a learning process throughout the life.
D31: Education is a continuous process that comes through experience.
D33: Education is an active and dynamic process.
D34: Person goes on reconstructing experiences throughout the whole life.

We have repeated the same procedure for three other domains and set of documents taken were of 50 count related to respective domains. We found the performance analysis of our method was much better as the result-set produced by it were having more meaningful pages.

Table 4: Ranking of the documents Relevance to User Query

SNo	Actual Rank	Google Rank	Variance by Google Rank	Our Rank	Variance by Our Rank
1.	D1, D2, D4, D3	D2, D4,D3, D1	10	D1, D2, D3, D4	2
2.	D21, D23, D25, D26, D22, D24	D21, D22, D23, D24, D25, D26	34	D21, D25, D26, D22, D23, D24	10
3.	D32, D33, D31, D34	D32, D33, D34, D31	18	D31, D32, D33, D34	6

IV. Performance Analysis

Performance of our approach for finding semantic similarity between the semantic web documents and the query considers not only on the keywords but also on the associated concept relations that exists between the concepts that are extracted from the document. This gives more specific similarity of the document with the user query. Then the spreading process for the query is also used for mapping using the statically model in which the document vector space model with the domain specific dictionary having weighted words. Along with this the ontology processor is quiet efficient to give the relevance value to the document with respect to the query.

We have compared the performance of our semantic similarity ranking scheme with the similarity computed using keyword based approach. We have taken around 50 documents related to three different domain. The processing of the documents using natural language processing techniques is done using earlier approaches and also the proposed novel ranking scheme is applied to these documents. We have tabulated the ranking of retrieved web pages by these techniques, to observe the effectiveness of our approach. We have observed that when given the query to the user interface it has been found that by using our approach the ranked list of the documents i.e result-set produced is having more meaningful pages accordingly to the query and having only few irrelevant pages which are of not of the user interest. After the analysis we found that our approach gives much better similarity measurement. The results obtained from the novel approach and have been presented in Table1.

We have analyzed the performance of our approach for deep analysis by finding the correlation with lexical matching and, we further looked to the pages retrieved from Google search engine having similar content but not repress the human view point. The large number of documents which were parsed having similar content were analyzed deeply with the help of our approach and ranked accordingly by taking the user view in mind which is usually not taken into consideration while ranking of the documents. We found that for maximum number of documents the approach produces good similarity measures with respect to the query of the user. In each case the similarity computation of our ranking model is much better than traditional ranking schemes showing the superiority.

V. Conclusion and Future Scope

The Semantic web provides several ways for improving search strategies and thus retrieving relevant web pages efficiently. The semantic similarity between the semantic web documents and the user query further improves the ranking by retrieving relevant web pages in the result-set produced by the search engine.

The novel ranking model presented in the paper takes the concepts and relationship between the concepts which exists both in the document and the user query to improve the retrieval of relevant documents in the result-set produced by the search engine. Our future efforts would be to design more meaningful and exhaustive ranking strategy by using the semantic analysis of web pages and by deeply statistical analysis relevance of documents, so that the semantic search engine can evaluate more precisely relevance and also the similarity between the web page and the user query. The ranking can even be done by any ontology already created or automatically creating a new ontology for the documents and the user query and then comparing them for the relevance score. We will also try to make our approach scalable for the semantic web.

REFERENCES

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

[2] Bollegala D., Matsuo Y., and Mitsuru, “A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web’, Proc of Int’l Conf on Empirical Methods in Natural Language Processing, pp 803-812, August 2009.

[3] Bollegala D., Matsuo Y., and Mitsuru, “A Web Search Engine-Based approach to measure Semantic Similarity between Words”, IEEE Transactions on knowledge and Data Engineering, vol. 23, no. 7, pp 977-990,July 2011.

[4] Cosma G. and Mike, “An approach to source-code Plagiarism Detection and Investigation using Latent Semantic Analysis”, IEEE Transaction on Computers, vol. 61, no. 3, pp 379-394, March 2012.

[5] 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.

[6] E. Greengrass, “Information Retrieval: A survey”. DOD Technical

[7] Report TR-R52-008-001, November 2000.

[8] Grossman D., and O. Frieder. “Information retrieval algorithms and heuristics”. Second ed. . Springer. 2004.

[9] Iosif E., and Potamianous, “Unsupervised Semantic Similarity Computation Between Terms usingWeb Documents”, IEEE Transaction on Knowledge and Data Engineering,vol. 22, no. 11, pp 1637-1647, November 2010.

[10] Lempel R., S. Moran. “The stochastic approach for link-structure analysis (SALSA) and the TKC e®ect”. In The Ninth International WWW Conference, May 2000.

[11] Li Z., and Karthik R., “Ontology-based Design Information Extraction and Retrieval”, Artificial Intelligence for Engineering Design, Analysis and Manufacturing, vol 21, pp 137-154, 2007.

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

[13] 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.

[14] Protiti M., “Semantic web: The future of WWW”, Proc. Of 5th Int’l Conf. CALIBER, Punjab University, Chandigarh, 08-10, 2007.

[15] Shamsfard M., Namehtzadeh A., and S. Motiee “ORank: An Ontology based System for Ranking Documents”, Int’l Journal of Computer Science, vol 1, no 3, ISSN 1306-4428, 2006.

[16] Tho Q., Hui S., and Tru C., ”Automatic Fuzzy Ontology Generation for Semantic Web”, IEEE Transactions on Knowledge and Data Engineering, Vol. 18., No. 6., June 2006.

[17] Vallet D., Fernández M., P. Castells, “An Ontology-Based information retrieval model”. 2nd European Semantic Web Conference (ESWC 2005). Heraklion, Greece, May 2005. Springer Verlag Lecture Notes in Computer Science, Volume 3532. Gómez-

[18] Pérez,A.;Euzenat,J.(Eds.),2005, Pages:455-470.

[19] 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.

[20] Zhong J., Zhu H., and Yong Y., “Conceptual graph Matching for Semantic Search”, Proc. of the 10th Int’l Conf. on Conceptual Structures: Integration and Interfaces pp 92-196 Springer-Verlag London, UK ©2002.