International Journal of Innovative Research in Engineering & Management (IJIREM)
ISSN: 2350-0557, Volume-3, Issue-6, November-2016
DOI: 10.21276/ijirem.2016.3.6.9

Semantic Based Re-ranking Model Using WordNet for Retrieving Web URL in E-learning

Subashka Ramesh
Saveetha University,
Chennai, India

Dr.A.Chandrasekar
St.Joseph College of Engineering,
Chennai, India

ABSTRACT

E-learning is fast, appropriate and just-in-time learning growth from the learning requirements of the new dynamically changing, distributed business world. Nowadays modeling user's preferences is one of the tricky tasks in e-learning systems that deal with huge volumes of information. The term "Semantic Web" encompasses efforts to build a new search engine that supports content with formal semantics, which Permits better potentialities for browsing and exploring through the cyberspace. Information retrieval by searching users query on the web is not a fresh idea but has different challenges when it compared to general information retrieval. Different search engine return different search results due to variation in indexing and search process. Semantic web can solve the problem in web with semantic annotations to provide intelligent and meaningful understanding by way of making use of query interface mechanism. In this work, we present a method for using genealogical data from ontology in finding the compatible hierarchical principles for question extension, and ranking websites founded on semantic family members of the hierarchical principles related to question terms, thinking of the hierarchical members of the family of domain searched (sibling, synonyms and hyponyms) via extraordinary weighting based on Re-ranking method. So, it provides an accurate answer for ranking records when compared to the previous methods.

Keywords: Semantic Web, Re-ranking, E-Learning, Tokenizer, Stopword Elimination, Parser

1. INTRODUCTION

Web based information retrieval systems, including Yahoo, Google and Bing recover mostly based on the basic tools to aid users to find textual data related with the World Wide Web. In spite of the vital role in reaching information, many of the retrieved results are irrelevant to the user's needs as they are ranked based on the string matching of the user's query. This has created a semantic gap between the meaning of the keywords in the retrieved documents and the meanings of the terms used in users' queries. However they filter the pages from searching unnecessary pages by utilizing evolved algorithms. These internet services can reply topic wise queries efficiently and effectively by developing state-of art algorithms. However, these systems don't take into account the semantic relationships between query terms and other concepts that might be significant to users.

If a user gives a query like "apple", Search engine retrieves the information about apple fruit and Apple Company (i.e) www.apple.co.in, iPhones, iPods, etc. But semantic search engine retrieves the document concerning apple fruit and its properties, relations etc. It retrieves the documents which are related to the user query. Thus, the addition of explicit semantics can improve the precision of the information retrieval systems using hyper graph re-ranking model and the proposed method will refine the search result from the web and improve the search process based on Information Retrieval technology using data from the Semantic Web [1]. This approach improves the traditional search that focuses on word frequency by trying to understand hidden meanings in the retrieved documents and users' queries.

This paper tries to overcome the problems based on retrieving textual information via word's meanings, rather than the word's lexical forms. Word sense disambiguation (WSD) technique is used to identify the concepts in context against Word Net ontology. Then the Resource Definition Framework (RDF) is used to annotate the semantics. We also propose the semantic retrieval process to realize semantically equivalent terms in documents and query terms using Word Net by associating such terms using semantic similarity methods.

2. RELATED WORKS

D. Manas, C. Hasan, S. Debakar, and A. Khandakar [4] have introduced the technique based on crawling of country based financial data. The focused method yield good results and precision with the aid of limiting themselves to a restrained area. It also tries to predict a target URL by pointing high quality web page before fetching the page. Their focused crawler is made for collecting the financial data for specific country.

H. Debashis, S. Biswajit, and K. Amritesh [5] were proposed to calculate the unvisited URL score based on text relevancy. They focused to calculate the similarity score in Google search based on text similarity and topic keyword with the Relevancy score of its parent pages. Relevancy score is calculated based on vector space model.

Z. Zhou, H. Jiang, J. Ma, X. Yang [6] has explained the construction process of document representation model based on content information and query. At the initial stage of information retrieval, they adopted traditional vector space to represent documents. Then, the information of the query space can be introduced into the document representation model gradually, thus the document-representing vector space becomes the integration of document space and query space. This model improve the efficiency and reliability of the feature terms of document.

Gyanendra Kumar and A K Sharma [7] proposed page rank based on visits of links (VOL). Page ranking algorithm is implemented based on visits of links for search engine which works on the basic ranking algorithm of Google. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behaviors which reduces the search space to large scale, but it does not take into account the page content.

A. Eneko, A. Xabier, O. Arantxa [8] proposed the Word Net expansion system over a collection with minimal textual information. The incorporation of knowledge through the use of ontologies has been very successful in many systems. Explicitly, Word Net has been used with many works related to text categorization, information retrieval and image retrieval.

3. PROPOSED SEMANTIC SIMILARITY TECHNIQUE

The proposed Semantic Similarity technique enhances the Query Expansion (QE) process using Query Manipulation. The QE is a system in knowledge Retrieval which consists of selecting and adding terms to the users query to improve retrieval performance from semantic web [9, 10]. The overall architecture of the system is shown in Fig. 1. The system is fragmented into number of segments for higher understanding. The semantic Similarity technique consist of phases like parsing of input query, Semantic parser, extracting synonym words using word net, semantic search with expanded query and Document Ranker based on weight table.

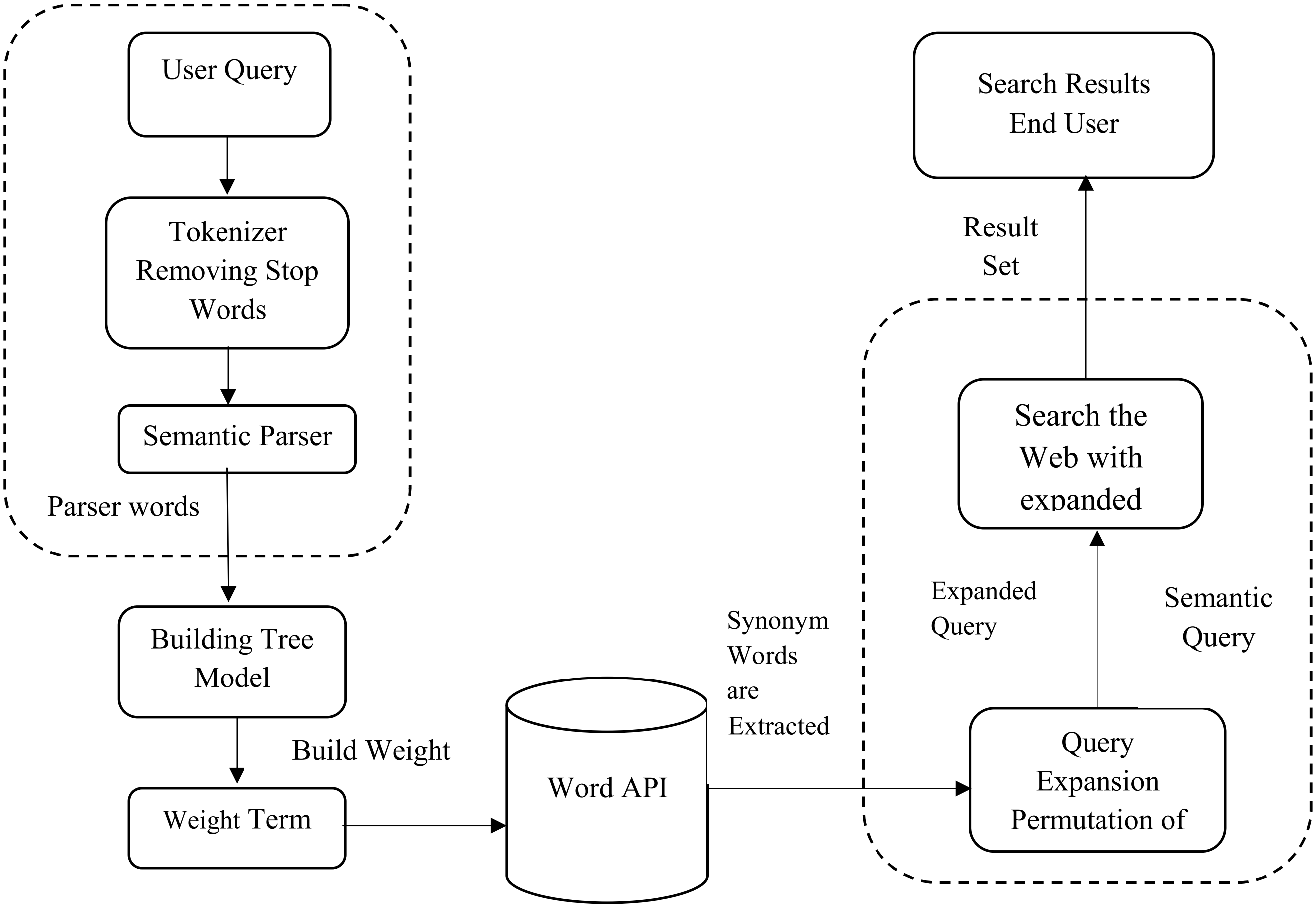

Fig.1 Architecture Diagram of Semantic Similarity Technique

The User enters input query to fetch data on E-Learning domain. The users query need to go through the semantic parser to analyze the given query. The parsing is done to analyze the query syntactically to determine every word in the query. The parsed output is saved in the file in the format shown as parsed sentence.

3.1 Algorithm of Semantic Parser

1. Parser splits the input queries into tokens using string tokenizer algorithm. The split keywords are called tokens.
2. Stopword elimination algorithm has been developed to eliminate Stopwords.
3. Stemming algorithm is developed to find the root phrase of the keyword. If the keyword is “searching”, it will reduce the word as “search”.
4. Processed keywords are called targets. The combination of Stopword elimination technique and stemming process are called modifiers.

5. Then, objectives are given as data to the WordNet. WordNet is a dictionary which is used to find synonyms of the targets. In this regard, the proposed approach is linked with the WordNet software.
6. Using the WordNet dictionary, the keywords are classified as either noun or verb or adjective. If the keyword is noun or verb, it is classified as subject. If it is adjective, it is labeled as predicate.
7. Apply the subject, predicate and object concept, RDF triples are formed.

Tokenizer

String Tokenizer allows a user query to break a string or text into tokens. A token is a categorized block of string consists of identifiers, numbers, quoted strings and indivisible characters.

Stopword Elimination

Stopwords are basically a set of common words that appear in the documents with little meaning; they provide only a syntactic operate but do not indicate area matter [11]. For example, in the context of search engine, if we Search question as “how you can develop know-how retrieval applications” and if the search engine tries to find web the pages that Include the phrases “how”, “to”, “boost”, “knowledge”, “retrieval”, “purposes” the search engine will find a lot of pages that contain The phrases how”, “to” than pages that incorporate the required information. So, if we ignore these two terms, the search engine can in reality focus on retrieving pages that include the keywords “develop”, “information”, “retrieval” ,”applications” which could closely bring up pages that are related to the users interest. The algorithm is implemented as given below.

1. The target record textual content is tokenized and individual words are stored in array.
2. A single Stopword is read from Stopword list.
3. The Stopword is compared to target text in form of array using sequential search technique.
4. It matches, the word in array is removed, and the comparison is sustained till length of array.
5. After removal of Stopword carefully, one more Stopword is study from Stopword record and again algorithm follows step 2. The algorithm runs constantly until all the Stopwords are compared.
6. Resultant text devoid of Stopwords is displayed, also required statistics like Stopword removed, number of Stopwords discarded from target text, total count of words in target text, count of words in resultant text, individual Stopword count found in target text is displayed.

4. RESULT FOR STOPWORD ELIMINATION

The implemented algorithm was tested with different collected document from web. Available 5 document text was feed into algorithm and the results obtained from system are listed in Table-1. Nearly 4600 Stopwords were removed thus reducing no. of words by approximately 13 %. Thus the efficiency obtained by the algorithm is approximately 97%.

Table-1 Stopword elimination details from documented text

S.no	Size in KB	Total Words in documents	No. of Stopwords eliminated	Percentage removal of Stopwords
1	157	5958	738	12.4
2	189	6033	952	15.8
3	210	9320	860	9.3
4	260	7045	912	12.9
5	324	8195	1145	13.9

Word net

WordNet is a lexical database of English. The database comprising nouns, verbs, adverbs and adjectives and organized by a collection of semantic relations into synonym units (synsets). These words are grouped and the results can be navigated and explored using web browser. In WordNet, two kinds of relations are used one is semantic and other is lexical[12]. The lexical relationships relationships hold between semantically related forms of words and the semantic relationships hold between related word definitions and these words are associated to form a hierarchy structure, which makes very useful Normal language processing. In this paper, we focal point on semantic similarity calculation captured by nouns and noun phrases. The four commonly used semantic relations are

- ➢ Hyponym/Hypernym (is-a)
- ➢ Section meronym/section holonym(part-of)
- ➢ Member meronym/member holonym(member-of)
- ➢ Substance meronym/substance holonym(substance-of)

For example, orange is a fruit (is-a) and Keyboard is part of computer (part-of) Hyponym/hypernym (is-a) is the most common relation used nearly 80% of the relations.

5. RERANKING ALGORITHM

Reranking plays an important role in searching. In this paper the documents are evaluated based on the semantic relation between Key terms in statements (semantic distance) and Key term frequency. The related Key terms found are weighted based on the result of the tree module. These two values are calculated according to the following subsections.

5.1Global/Local Key Term Extraction

In this Paper, We define two kinds of Key Terms. Global Key Terms obtain from the whole document set and Local Key Terms obtain from a single document or a query. We adopt a two-stage approach to automatically acquire Global Key Terms and Local Key Terms. After we have attained Global Key Terms and Local Key Terms, we make use of them to reorder the top M (M<=1000) documents in initial ranking documents. Suppose q is a query, Fig. 2 is the algorithm to reorder top M documents in initial ranking documents where w(t) is the weight assigned to Local Key Term t. w(t) can be assigned different value by different measures.

```
For example,
        w(t) = the length of t;
        w(t) = the number of Characters in t;
        w(t) = square root of the length of t;
        w(t) = square root of the number of  Characters in t;
(default)
        for each document d in top M ranking documents
         sim ← similary value between d and q;
        w ← 0;
        for each Local Key Term t in query q;
         {
              if t is a Local Key Term of d w ← w + weight(t)
        };
         if (w > 0)
        {
                   sim ← sim * w;
                    set sim as the new similary between d and q
         };
 reorder top M documents by their new similarity values with
query q;
```

6. COMPARATIVE ANALYSIS

The effectiveness of proposed system is analyzed by semantic technique using Reranking algorithm. The user query word is connected with the WordNet dictionary. With the meanings of the query, the documents are retrieved from the database. Table-2 shows the comparison between WordNet based Semantic technique and Non-WordNet based Search. A graph is drawn using the data given in Table-2. There are two measures used for evaluating the effectiveness of semantic search method. They are precision and bear in mind. Precision is the ratio of total number of retrieved documents to the total number of documents in the database. Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Precision} = \frac{\text{Total number of documents retrieved by a search engine}}{\text{Total no. of Documents}}$$

$$\text{Recall} = \frac{\text{Total number of relevant documents retrieved}}{\text{Total number of retrieved documents}}$$

Fig-2 Algorithm for Reranking document

Table-2 Comparison between Non WordNet and WordNet based Semantic Search

Type of Search	Total No of Documents in the database	Total No of Retrieved Documents for User Query	Total No of Relevant Documents in Retrieved Documents for User Query	Precision	Recall
Non WordNet based Semantic Search	150	65	48	0.44%	73.8%
WordNet Based Semantic Search using Reranking algorithm	150	92	83	0.62%	90.21%

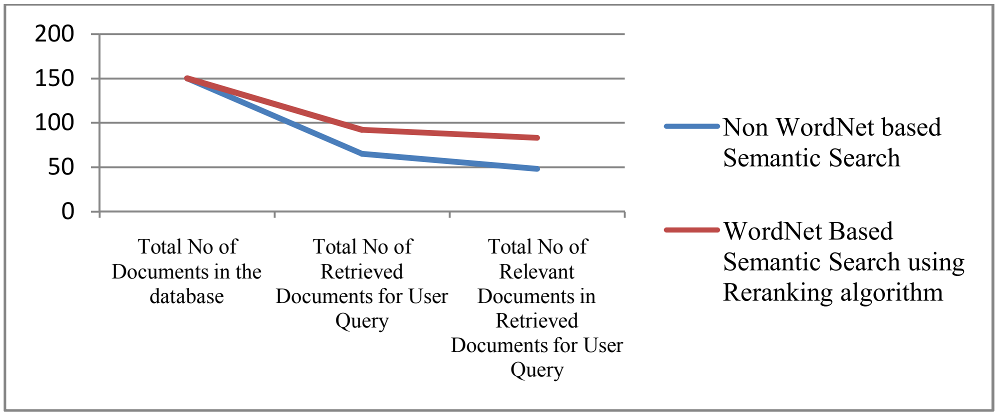

Figure- 3: Comparative Analysis using WordNet

7. CONCLUSION

In this paper, the semantic search is developed using the semantic concepts and the search process is improved to overcome the traditional search problems by semantic technique using Reranking algorithm. The Proposed system is compared with the Non WordNet based Search. The implementation results show the performance in the information retrieval. The performance of the system is measured by using precision and recall method.

8. FUTURE WORK

In this work we focus on single ontology for single domain. In future, we will focus on multiple ontologies which allow us to give an opportunity for employing the knowledge from different ontologies of single or different domains.

REFERENCES

[1] Preethi , Ms.N., and Devi , Dr.T., Case and Relation (CARE) based Page Rank Algorithm for Semantic Web Search Engines. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.

[2] Ch.-Qin Huan,Ru-Lin Duan, Y. Tang, Zhi-Ting Zhu, Y.-Jian Yan, and Yu-Qing Guo, ."EIIS: an educational information intelligent search engine supported by semantic services".International Journal of Distance Education Technologies ,January 1, 2011.

[3] Robin Sharma , Ankita Kandpa,and Priyanka Bhakuni, Rashmi Chauhan, R.H. Goudar and Asit Tyagi." Web Page Indexing through Page Ranking for Effective Semantic Search". Proceedings of 7th International Conference on Intelligent Systems and Control (ISCO 2013).

[4] D. Manas, C. Hasan, S. Debakar, A. Khandakar. (2010). Focused Web Crawling: A Framework for Crawling of Country Based Financial Data. 978-1-4244-6928-4/10, IEEE.

[5] H. Debashis, S. Biswajit, K. Amritesh. (2010). Adaptive Focused Crawling Based on Link Analysis. 201O 2nd International Conference on Education Technology and Computer (ICETC). 978-1-4244-6370-1, IEEE.

[6] Z. Zhou, H. Jiang, J. Ma, X. Yang (2010). Study on Application of Document Representation Model Based on Query and Content Information in Website Search Engine. 2010 International Conference on Web Information Systems and Mining, 978-0-7695-4224-9/10, IEEE.

[7] Gyanendra Kumar et al., “Page Ranking Based on Number of Visits of Links of Webpage”, YMCA University of Science & Technology, Faridabad, India(ICCCT-2011)

[8] A. Eneko, A. Xabier, O. Arantxa. (2010). Document Expansion Based on WordNet for Robust IR. COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics, Volume, pages 9–17, Beijing. ACM.

[9] Preethi , Ms.N., and Devi , Dr.T., Case and Relation (CARE) based Page Rank Algorithm for Semantic Web Search Engines. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.

[10] Lee, T. B., Hendler, J., and Lassila ,O.,"The semantic web". Scientific American, vol. 284(5), May 2001.

[11] Riyad A, Ghassan K, Jihad J, Ahmad H and Eyad H, “Stop-Word Removal Algorithm for Arabic Language”, Information and Communication Technologies: From theory to Applications, 2004 proceedings, IEEE 2004.

[12] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis and E. E. Milios, “Semantic similarity methods in WordNet and their application to information retrieval on the web”, Proceedings of the 7th annual ACM international workshop on Web information and data management, (2005) October 31-November 05, Bremen, Germany.