Semantic Web Search Engine

Amisha Mehta[1], Pranav Makkar[2] , Saylee Palande[3], Prof. S. B Wankhede[4]
Department of Computer Engineering
Rajiv Gandhi Institute of Technology.
Versova, Andheri(W),
Mumbai-400053.

Abstract:- **The World Wide Web (WWW) allows people to share information or data from the large database repositories globally. We need to search the information with specialized tools known generically as search engines. There are many search engines available today, where retrieving meaningful information is difficult. However to overcome this problem of retrieving meaningful information intelligently in common search engines, semantic web technologies are playing a major role. In this paper we present a different implementation of semantic search engine and the role of semantic relatedness to provide relevant results. The concept of Semantic Relatedness is connected with Wordnet which is a lexical database of words. We also made use of TF-IDF algorithm to calculate word frequency in each and every webpage and Keyword Extraction in order to extract only useful keywords from a huge set of words. These algorithms are used to retrieve much optimized and useful results to the user.**

Keywords: TF-IDF, Semantic Relatedness, Keyword Extraction, Wordnet

1. INTRODUCTION:

Most of the traditional search engines search for keywords to answer queries from users. The main focus of these search engines is solving queries with close to precise results in small period of time using much advanced algorithms. However, it shows that such search engines are incompetent in answering queries intelligently using traditional approach. This is where semantic web search engine comes into picture. The Semantic Web will support more efficient discovery, automation, integration and reuse of data and provide support for interoperability problem which cannot be resolved with current web technologies. In short it will intelligently understand the user query and search for those results that match not only the keyword but also the meaning of that query. In this paper, we will make modification over the existing search engine by adding an additional concept of keyword extraction and semantic relatedness calculation. Semantic relatedness here, is a metric which calculates the relation between words. This metric is computed with the help of Wordnet. Another metric used in the current approach is TF-IDF (Term Frequency-Inverse Document Frequency). It is used to calculate the relevancy of each word and relevance of each document.

2. LITERATURE REVIEW:

2.1 Literature review on SWSE based on Ontologies and concepts:

Cordi V[6] proposed an ontology-based similarity between sets of concepts in order to evaluate the information present in one document with other document with respect to ontology. Here the concepts are extracted and semantic similarity is computed between them by using a modified Djikshtra's Algorithm of the graph theory.

Pisharody A[8] proposed a search engine based on keyword relations. In this, creation of a database is done, which consists of words and their relations to keywords that overcome the drawback of keyword based approach. The LGP (Link Grammar Parser) is used to parse the web pages and normalization process is used to remove the duplicate values and remaining nouns, adjectives and verbs are stored in the database. These words are fed into Wordnet to determine the set of relations. When a query is given by the user, it undergoes the normalization process and each query word is searched in the database and if the word is not present then Reverse Lookup algorithm is executed.

Thiagarajan R[7] proposed evaluating semantic similarity using ontologies. In this paper, Web pages can be represented in a Bag of Concepts (BOC). In BOC, there are concepts which semantically represent the web page. For computing semantic similarity between web pages, spreading is used. Spreading means including additional related terms to an entity by referring to ontology such as Wordnet and Wikipedia.

Li Y.[9] proposed relation based search engine. The authors proposed a search engine ONTOLOOK which considers the relations between concepts from a web page. In this search engine process, analysing of input keywords is done. Later the keywords are paired with some concepts and these pairs are sent to the ontology database to retrieve all the relations. A concept-relation graph is formed based on these retrieved relations and concepts. This graph is trimmed and constructs sub-graphs which will be used to form property-keyword candidate set based on corresponding keyword pairs and then it is sent to the database. The authors use Google PageRank ranking algorithm.

In the work surveyed, the main focus is on introducing semantics either by taking ontology or relationship that exists between the concepts.

2.2 Literature review on Semantic Relatedness Measures :

The Calculation of Semantic Relatedness can be done with the help of an important tool called "Wordnet". With the help this lexical tool many algorithms are proposed in order to calculate semantic relatedness. We can distinguish in three ways to determine the semantic similarity between objects in ontology: The first approach indicates the evaluation of similarity by information content (also called the IC/node based approach). The second approach represents an evaluation of similarity based on conceptual distance (also called path/edge based approach). The third approach is hybrid and combines the first two approaches. This relatedness approach focuses on words and ignores ontological relationship.

Author and Year of publication	Description	Approach
Wu & Palmer [5],1994	This measure considers the depth of the synsets in the Wordnet taxonomies and calculates relatedness.	Path Based
Resnik [4],1995	Resnik calculate and returns the information content of the lowest common subsumer(LCS).	Node Based
Jaing & Conrath [1],1997	This measure combines edge counts and information content values in the Wordnet.	Node Based
Lin [3],1998	This Measure divides the information content of the LCS with the sum of the information content of the two synsets.	Node Based
Leacock & Chodorow [2], 1998	This measure depends upon the length of the shortest path between two synsets and this measure limit their attention to IS-A links.	Path Based

Table No.1 Literature Survey on Semantic relatedness

3. THEORETICAL FOUNDATION OF PROPOSED APPROACH:

We present in this section the theoretical basis of our proposal. These features guide the semantic evaluation approach that we propose. For this purpose, we are making use of test collection of web pages, that is, by making use of a dataset. This dataset consists of 25 pages. This system is based partly on a linguistic resource Wordnet. Semantic Evaluation is done for every word in each web page. This Semantic Evaluation is called the Semantic Relatedness of a word. Our system also does extraction of important keywords from all the Web pages. Extraction of Keywords from Web pages is done with the help of a tool called AlchemyAPI. Based on the query entered by the user the TF-IDF of each page is calculated. Traditional Search Engine does only frequency calculation, which means it only calculates the frequency of the query entered by the user in the pages; whereas our Semantic Web Search Engine does the following also:

i. It extracts important keywords from the web pages.

ii. It calculates Term Frequency and Inverse Document Frequency and assigns every document a fix value.

iii. It then calculates the Semantic Relatedness of Each Page.

iv. It sorts the pages on the basis of TF-IDF score plus Semantic Relatedness Score.

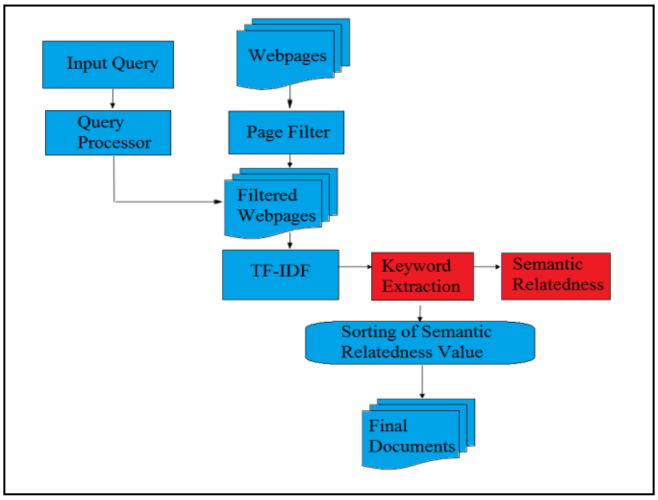


Fig 1. Architecture of Semantic Search Engine.

More details of the proposed approach are as follows:

3.1 TF-IDF:

The term weighting function known as TF-IDF was proposed in 1972, and has since been extremely widely used. The Information Theory Approaches of this algorithm are problematic, but that there are good theoretical justifications of TF-IDF algorithm in traditional probabilistic model of information retrieval. In this algorithm term frequency which is also called as TF measures the number of times a term occurs on a page. One problem here is that each document has different size. On large pages the frequency of the term is higher and on small pages the frequency of the term is lower. So in such a case we are trying to normalize the page based on its size. A simple solution is to divide the term frequency by the total number of terms in a page. For example, if a page consist of a word "dog" 2 times and if the total number of word in that page is 10, then the TF value is 2/10=0.2.

Similarly, the IDF value is also calculated. Inverse Document Frequency also called as IDF value is based on the number of pages in the collection being search which contains the words which are entered by the user. The primary purpose of searching is to find the relevant documents matching the query. In TF calculation all terms are considered equally important. In fact, certain terms that occur too frequently were considered less important, like "is", "the", "and", "of", "but" etc. Therefore we need to weigh the value of less frequent words and increase the importance of the page in which occurrence of the term is less frequent. The relevancy of such pages should be higher. For this purpose we make use of logarithms.

Suppose a term "adventure" is being searched.

IDF value is given as: 1+log (Total Number of Pages/Number of pages containing the term "adventure").

In order to give each document a unique value we multiply the TF value with the IDF value. This algorithm is the first step of Information Retrieval in our Search Engine.

3.3 *SEMANTIC* RELATEDNESS:

Semantic relatedness measures can be used for performing tasks such as term disambiguation, as well as text segmentation. In comparison to the above stated measures in Section 2 (Literature Review), we see that the measure given by Wu and Palmer is simple, and gives good performance. That's why we are using this similarity measure for calculating semantic relatedness. Nevertheless, the Wu and Palmer measure has a disadvantage: in some situations, the similarity of two elements of an IS-An ontology contained in the neighbourhood exceeds the similarity value of two elements contained in the same hierarchy[5].

3.3.1 Wu and Palmer Measure:

The basis of similarity computation is based on the edge counting method, defined as follows:

Given an ontology Ω formed by a set of nodes and a root node (R) as shown in Figure 2, we will calculate similarity of two ontology elements C1 and C2. The basis of similarity computation is based on the distance (D1 and D2) which separates nodes C1 and C2 from R and distance (D) separates the lowest common subsumer(CS) of C1 and C2 from R. The similarity measure of Wu & Palmer[5] is defined by the following expression:

$$SimWP = \frac{2*D}{(D1+D2)} \qquad (1)$$

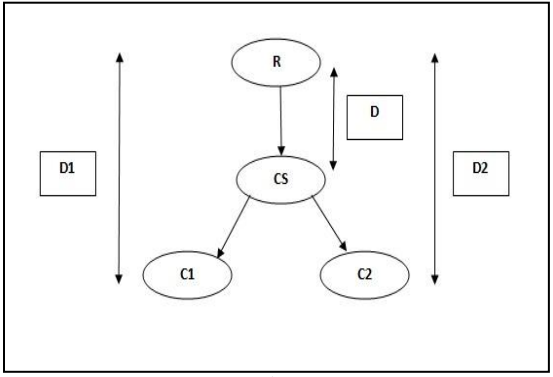


Fig 2. Example of a concept hierarchy

In Figure 3, we present a graph representing a hierarchy of the concept. This graph represents an

Ontology extract of pedagogic field.

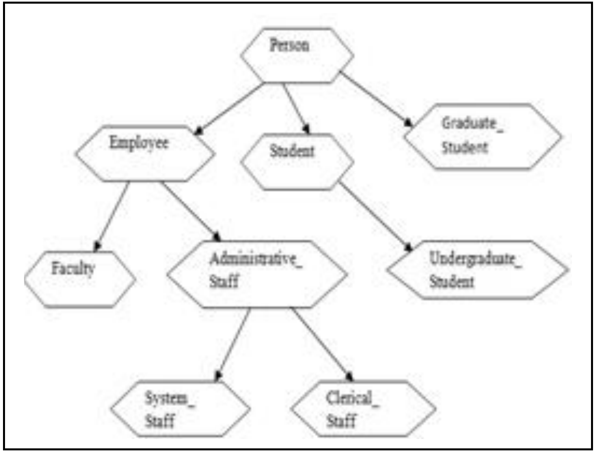


Fig 3. Concept Hierarchy

Let in the ontology of Figure 3, C1, C2 and C3 be the concepts "Person", "PostDoc" and "Administrative_Staff". SimWP (C1, C2) =2*1/ (1+4) =0.4 and SimWP (C2, C3) =2*2/ (4+3) =4/7=0.57

As we can see, similarity values obtained by this measure show that neighbour concepts C2 and C3 are more similar than the concepts C1 and C2 located in same hierarchy. This is problematic and inefficient. Therefore, a new modified measure has been introduced which inherits the advantages of Wu and Palmer work and represented by the following formula [10]:

$$\mathrm{SimWP(C1, C2)} = \frac{2*\mathrm{D}}{(\mathrm{D1+D2})} * \mathrm{PF(C1, C2)} \qquad (2)$$

Let PF(C1, C2) be the penalization factor of two concepts C1 and C2 and it is defined as in (3):

$$\mathrm{PF(C1, C2)} = (1 - \lambda).(\min(\mathrm{D1, D2}) - D) + \lambda\,(|\mathrm{N1} - \mathrm{N2}| + 1)^{-1} \qquad (3)$$

Let D1 and D2 be the distances which separate nodes C1 and C2 from the root node, and D, the distance which separates the lowest common subsumer of C1 and C2 from the root node. The coefficient λ is a Boolean value indicating 0 or 1, with 0 indicating two concepts in the same hierarchy and 1 indicating two concepts in neighbourhood.

3.4 KEYWORD EXTRACTION:

Calculating relatedness of all the words in a web page is inefficient. Stop words, which are the frequently occurring words, are removed from the Webpage. For example, “the”, “is”, “hence”, “therefore”, “we”, “you”, etc. Removal of stop words makes the search fast and precise. Apart from removal of stop words, important keywords from all the web pages are extracted with the help of a tool called Alchemy API. Alchemy API uses natural language processing techniques and machine learning technologies to extract semantic meta-data from the web page content. Information related to people, relationships, authors, topics, facts are extracted from the webpage. We are extracting verbs and nouns from all the web pages. This Keyword extraction algorithm is used by using end points of Alchemy API. Keyword Extraction reduces the number of terms to such a great extent that calculating the Semantic Relatedness of each word becomes very fast. Alchemy API also indexes important keywords and ranks them. This API works on URLs, HTML documents, plain text. The API gives an XML format of the webpage. This format consists of relevance number on the basis of which we extracted more useful keywords.

3.5 Wordnet:

Wordnet is an Electronic Dictionary of nouns, verbs, adjective and adverbs developed which is developed at Princeton University. It organizes related set of words in a Synonym set also called as Synsets. For example {automobile, car, bike, machine}. All the words in Wordnet are stored as a tree like structure. In order to calculate Semantic Relatedness of words we are making use of Wu and Palmer Measure as mentioned above. This Wu and Palmer measure make use of Wordnet. It calculates relatedness by considering the depths of the synsets in the

Wordnet. Every word in the WORDNET has different sense. For example, sense of the word "Dog" is,

i. Someone who is morally reprehensible.
ii. Animal Etc.

Wu Palmer Algorithm in short checks the sense of each word which in Wordnet is stored as a tree like structure and it compares the glosses of each word sense. The gloss with maximum keyword matches is said to have the highest relatedness value. In Wu Palmer the relatedness value of '1' is said to be the highest value and the value 0.0 is said to be the lowest relatedness value. Every word in the WORDNET is taken as "Concept" in Wu and Palmer Algorithm.

4. RESULTS:

In this Search Engine we have used JSOUP Library which is a Java Library which is used to parse HTML and XML pages into normal text format. We made used of a mini dataset of 25 pages related to food and animals. The Query entered by the user is also dataset specific. The Query Entered by the user is pre-processed at the back end. This means that the Stop words from the query is eliminated and additional stemming is done to each word. For example, if user enters "dangerous", the word will get converted to danger. Here is the Snapshot of our project:

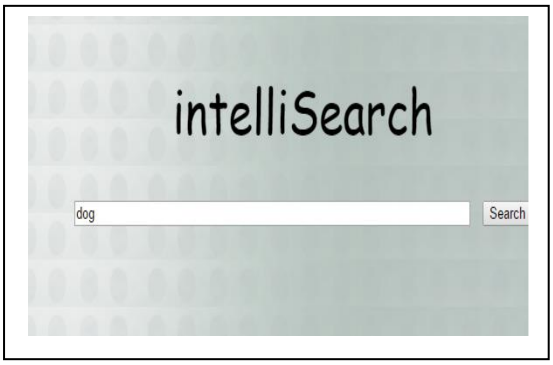
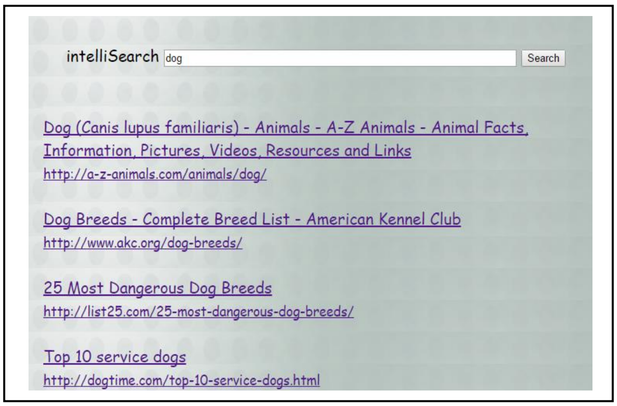


Fig 4. Entering of a query "dog"



Dog (Canis lupus familiaris) - Animals - A-Z Animals - Animal Facts, Information, Pictures, Videos, Resources and Links
http://a-z-animals.com/animals/dog/

Dog Breeds - Complete Breed List - American Kennel Club
http://www.akc.org/dog-breeds/

25 Most Dangerous Dog Breeds
http://list25.com/25-most-dangerous-dog-breeds/

Top 10 service dogs
http://dogtime.com/top-10-service-dogs.html

Fig 5. Initial 4 search results.

These 4 links displayed in Figure 5. Are linked which is retrieved after the keywords extraction and semantic relatedness. This means that every page displayed has its own unique value. These unique values are then sorted with the help of sorting algorithm in descending order and then the results are displayed.

The difference in the value obtained is shown in Table 2 below:

Query	Term Frequency and Inverse Term Frequency		SR Score
	Document URL	TF-IDF value	
Dog	http://dogtime.com/dog-breeds	0.02	0.42

Table 2. Results for TF-IDF and Semantic Relatedness for the word 'Dog'.

Traditional Search engines often search only for the frequency of the Query word in web pages but the one thing which is forgotten is that every word has some relation with the other word. For example, the word 'dog' has high relation with the word 'animal'. This is the main purpose of Semantic Relatedness.

5. CONCLUSION AND FUTURE SCOPE:

This Semantic Web Search Engine has improved the search quality to great extent. As the traditional search engine only checks for the frequency of the word; this Search Engine looks for the relation that is the relatedness of the words in a web page. It has also improved efficiency due to the keyword extraction algorithm. Due to this algorithm the total number of words on which the semantic relatedness has to be applied is reduced to a greater extent. Our system could be improved in future by implementing multithreading in order to improve the speed of the search engine and by eliminating the dependency of API. The scope of the project can be extended to making use of large dataset containing more number of documents if the use of multithreading is done successfully. Implementation of word sense disambiguation will improve the performance and will give the users the most useful pages which can also be added as future work.

6. REFRENCES:

[1] Jiang J. and Conrath D. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings of International Conference on Research in Computational Linguistics, Taiwan.

[2] Leacock C. and Chodorow M. 1998. Combining local context and Wordnet similarity for word sense identification. In Fellbaum 1998, pp. 265-283.

[3] Lin D. 1998. An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI.

[4] Resnik P. 1995. Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448-453, Montreal.

[5] Wu Z. and Palmer M. 1994. Verb Semantics and Lexical Selection. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Las Cruces, New Mexico.

[6] Cordi V, Lombardi M, Viviana M, "An ontology-based similarity between sets of concepts", WOA, 2005.

[7] Thiagarajan R., Manjunath G., and Markus S. (2008), Computing semantic similarity using ontologies ISWC 08, the International Semantic Web Conference (ISWC), Karlsruhe,Germany.

[8] 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.

[9] Yufei Li, Yuan Wang, and Xiaotao Huang,"A Relation-Based Search Engine in Semantic Web",IEEE transactions on knowledge and data engineering,Vol. 19,No. 2,February 2007.

[10] Vadivu Ganesan, Rajendran Swaminathan, M.Thenmozhi, "Similarity Measure Based On Edge Counting Using Ontology",International Journal of Engineering Research and Development,Volume 3, Issue 3 (August 2012).