SEMANTIC TERM BASED INFORMATION RETRIEVAL USING ONTOLOGY

J. Mannar Mannan[1] Dr. M Sundarambal[2]

1. Department of Information Technology, Anna University, Regional Centre Coimbatore.
2. Department of Electrical and Electronics Engineering, Coimbatore Institute of Technology, Coimbatore.
E-mail:endeavour6381@yahoo.co.in

***Abstract*: Information Searching and retrieval is a challenging task in the traditional keyword based textual information retrieval system. In the growing information age, adding huge data every day the searching problem also augmented. Keyword based retrieval system returns bulk of junk document irrelevant to query. To address the limitations, this paper proposed query terms along with semantic terms for information retrieval using multiple ontology reference. User query sometimes reflects multiple domain of interest that persist us to collect semantically related ontologies. If no related ontology exists then WordNet ontology used to retrieve semantic terms related to query term. In this approach, classes on the ontology derived as semantic related text keywords, these keywords considered for rank the documents.**

Keywords – Information Retrieval, Semantic Term, WordNet, Ontology.

I. INTRODUCTION

The challenging process on internet is Information retrieval (IR). It is a process of analyzing, retrieving and clustering of information based on domain of interests. The target motive of IR is to provide relevant document to the given query. Many IR systems in early days suggest keyword based retrieval that the exact keyword found in the documents. Words in a document are roughly classified into three ways (1) Special Characters, it is a character like, semicolon, punctuation, comma etc., (2) Stop words, word spread across the document do not provide any meanings such that 'the', 'a'. 'and' ,'then', etc., [3] Provides the methods for how to evaluate stop word, and [8] suggested problem causes by the stop words. (3) Keywords, used to rank the documents. Maximum of IR system working based on the frequency of keywords spread across the documents.

A. Basic Methods

The Boolean model is a simple model follows a set theory, based on the "exact-match" principle. A drawback with the model is its discreteness, not allowing any relevance ranking. Latent Semantic Indexing (LSI) is a index based retrieval method uses Singular Value Decomposition (SVD) that provides reduced model for represent term-to-term, document-to-document and term-to-document matrix. By reducing dimension, it is possible for document with deferent profiles of terms usage to be mapped into the same vector of factor values. This helps to eliminate the noise in the original data, thus improving the reliability of the algorithm. But, all these methods working on query terms and fail to considering the semantic term. In the new information age, keyword based retrieval not enough to implement on huge amount document like internet. To address this issue, ontology is utilized for information integration. Ontology is referred as domain of knowledge, represents the concepts and relationship among the concepts.

B. Ontology

Ontology is concept that describes the resource and different kinds of relationship among the resources. It is a new trend applicable in the field of artificial intelligence and modern information retrieval that describe meaningful relation among things (resources). The ontology described in triples such that resource as 'classes', relation as 'properties' and 'instances'. Semantic IR system make use of these triples for identify the relationship between two or more keyword for retrieval. Many ontology development tools exist to develop ontology like protégé, OntoEdit etc., and build-in ontologies are also available like WordNet, Cyc and DBpedia.

C. WordNet Lexical analyser

WordNet is developed by Princeton University, is an elixir for information retrieval, which is used in various purposes. It is a lexical analyser used in natural language processing (English) contains around 150000 synsets and their semantic relations. Many open source java API for WordNet available in internet. edu.sussex.nlp.jws.jar; and edu.mit.jwi.jar; is a Java API providing interface to WordNet to retrieve information using application programme. Similarly, biggest ontology so far is DBpedia contains around 2.4 million resources and Cyc ontology contains around 300000 concepts. These in build ontologies ease our work in

information retrieval, but make time consumable when ontology grows rapidly. The authors [7] utilized WordNet for adaptive search using structured representation of data.

II. RELATED WORK

From the given set of words as query to IR system, the syntactic and semantic taxonomy carried out to measure the similarity between words to guess weighty of the document corresponding to query. The Explicit Semantic Analysis (ESA) is a concept based algorithm [1] that maps the concepts and query using Wikipedia-ESA concepts space and perform indexing for information retrieval. This model suffers due to ambiguity in timing data and retrieves only static documents. Consider this model may fail to integrate temporal data such that temperature of New Delhi. As in [2] similarity between two words 'W1' and 'W2' measured by Sim (W1,W2); if 'W1' and 'W2' are semantically equal, then its value closer to 1 and if not then closer to 0. From this measure, the authors relate the keyword for information retrieval. From the authors point of view, the term 'computer' and 'network' relation value is 0.7368.(WordNet). But semantically both the terms have high level relationship. If the threshold value between words set to 0.9 then the term 'computer' and 'network' not semantically related. From this suggestion the IR system fail to analysis semantic terms. In this proposed method ontologies that highly related to user query stored in the local repository are selected, than using WordNet ontology finds the relation between one or more ontologies. If there is no appropriate ontology in local repository, than the WordNet used to extend the query term to get the semantic terms. The authors [4] suggest taking annotation for any given query to strengthen the process of document ranking measurement.

Ontology selection needs two basic groundwork such that tokenization and stemming. Tokenization is processes of split the line or paragraph into individual keywords to compute the frequency of occurrences of keywords. Stemming is a process of identify the route word of a given keyword. As in [6] provide the meaning full way of the roll of WordNet in matching two or more ontology by analyzing its classes and properties.

Ontology plays vital role in IR and sometime more than one ontologies collectively involved in complex information processing. It is directly possible for retrieval the semantic terms with WordNet; it returns the value for 'thing' and 'object' is 0.9473. From the human understanding, it is true. But, it also returns the value for 'Missile' and 'Rocket' is 0.9523. From this point, it is not true. Both 'missile' and 'rocket' may have some common property, but technologically both are totally different. From the observation, WordNet sometime not trusted for information integration. Here, the ontologies in local repository are used for semantic term analysis. As in [5] provide ontology mapping techniques to integrate more ontology to map common domains, by its nature of enlarging, not suitable for IR. As in [9] provided the secured key term for kids that are collected across many schools. It is an approach provides restriction on accessing contents of internet by the kids. The author [10] used the database with trained collection of documents. In this approach, the authors never consider the semantic terms that highly related to query.

III SEMANTIC TERM BASED INFORMATION RETRIEVAL

A. User Query Processing

Query processing is the first attempt of any IR system to identifying the requirements in terms literal words. It is a process of understanding user intended domain of interest. Sometime query reflect multiple domain of interest. After removing strop words and other unnecessary terms, remaining terms are treated as conceptual keywords. The authors [11] attempted for query modification by semantic term relation between queries by concept mapping. Author compare logs of two search engines for query modification. Comparing two search engines logs is a time consuming process. In our proposed method, the keywords are directly compared with keywords of ontologies that are converted into text document found in the local repository. If there is no appropriate ontology found in the local repository, the WordNet ontology is utilized to expand the keywords. From this process, keywords with set of related semantic terms are grouped to rank the documents. Figure 1 shows the architecture diagram for semantic term based information retrieval.

B. Domain Ontology in Local Repository

Ontology that describes the specific domain of real world thing is domain ontology. To exemplify 'computer' is a domain and ontology describes hardware, software, operating system, file system, memory, etc., This ontologies can be created using protégé, OntoEdit or download from semantic search engine "*swoogle.umbc.edu*". The selection of domain ontology for corresponding query terms is a critical process. The authors [12] provide new method to measure the similarity value for selection of domain ontology to the given query. The first process of selection is related the query term with classes of ontology found in the local repository using WordNet. Comparing the query terms with all classes found in the ontology using WordNet, and if the relationship value between query term and classes ≥ 0.8 is considered for selection. The highest mean value of ontology is selected for domain ontology. As in [13] provides the semantic similarity measure by key word present between two sentences. It is a simply a keyword based matching technique, does not provide any semantic weight to the document.

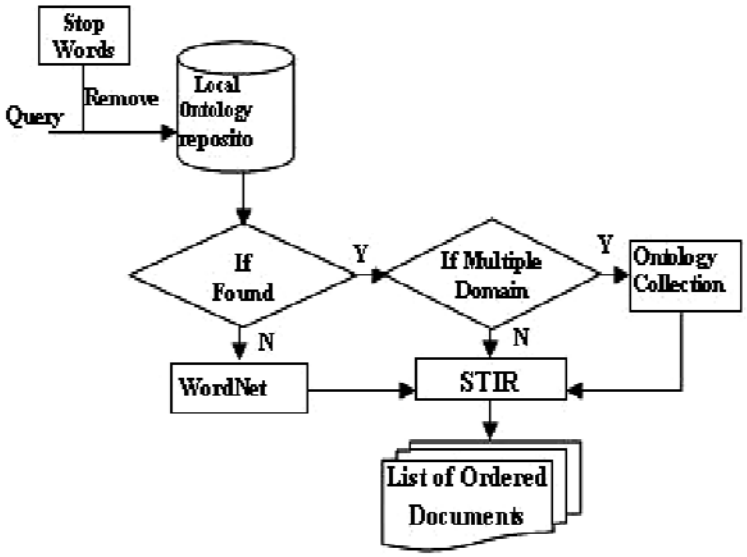

Fig.1. Architecture Diagram

C. Ontology Text Processing

After successful selection of domain ontology, it is derived as a text documents to collect the classes and treat as semantic keywords. The collections of semantic keywords in the ontology along with user keyword are considered for rank the documents in the proposed method. More number of classes in the ontology sometime misleads the document ranking.

```
Algorithm semweigh(keyterms ,Documents)
Begin
  For each document di in |D|;
   dl=word(di);
    For each term tj in document di;
      For each Semantic term sk for each term tj;
        If ((Domain Ontology)==true)//Found in local repository
              If(sim(tj,sk)≥8.0)
                  Semantic [d] += s[k];
                  Count++;
              endif
       else // using WordNet to get semantic terms
              s[k]=WordNet(tj);
             If(sim(tj,sk)≥8.0)
                  Semantic[d] += s[k];
                  Count++;
            endif
     End k;   End j; End i;
Semweight=count/dl;
End;
```

To filer the number of terms from the ontology, WordNet ontology used to measure the relationship value between key terms with domain term. If the value found suitable, then it will take as a semantic terms, if not, the terms simple neglected. The algorithm shows to measure similarity value.

To compute the mean value of the term frequency of a document corresponding to only query term is given in equation 1.

$$Wd_i t_j = \frac{d_i tf_j}{dl} * 100 \quad \text{.....(1)}$$

Where $Wd_i t_j$ - is the total weight of the document 'd_i' based on query term frequency 'tf_j'. The semantic terms can be measured in the same way shown in equation 2.

$$Wd_i st_j = \frac{d_i st_j}{dl} * 100 \quad \text{.....(2)}$$

The semantic terms related with the query terms retrieved from corresponding ontology and is excluding the query terms. The total weight of the document is measured by adding semantic term frequency with query term frequency.

$$TWd_i = Wd_i t_j + Wd_i st_j \quad \ldots\ldots(3)$$

The total weight of the page is measured by the mean value of concatenation of total term frequency associated with semantic term frequency spread in the document.

IV. EXPERIMENTAL EVALUATION

To validate our proposed model, 1000 word documents and the query 'communication hardware' taken for analysis. From the evaluation, 'computer' as basic concepts that user highly prefers to retrieve and 'network' is an additional weight added to the domain. From the WordNet ontology, Query Term (qt1) 'communication', provides 3 different sense, likewise 'qt_2', have five senses and more than ten meronyms.

The ten document collection d2, d2,…..d10 is considered for experimental evaluation out of 1000 experimented documents. The key term distribution of ten documents has been tabulated TABLE I.

TABLE I
Term Frequency Distribution

Query Term	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
Communication	6	2	1	2	2	3	5	6	1	3
Hardware	9	3	1	3	1	1	3	2	3	2
Total	**15**	**5**	**2**	**5**	**3**	**4**	**8**	**8**	**4**	**5**

After removal of stop words in a standard page, the average number of terms is considered as dl= 950 for evaluation. The corresponding average term frequency measured by $Wd_i t_j = \frac{d_i tf_j}{dl} * 100$ and the average term frequency of each page in Table I is tabulated in TABLE II.

TABLE II
Average Term Frequency

d_i	tf_j	$\frac{d_i tf_j}{dl}$
D1	15	0.015789474
D2	5	0.005263158
D3	2	0.002105263
D4	5	0.005263158
D5	3	0.003157895
D6	4	0.004210526
D7	8	0.008421053
D8	8	0.008421053
D9	4	0.004210526
D10	5	0.005263158

TABLE III
Semantic Term Frequency Distribution

Ontology	**Semantic Terms**	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
Ontology1	**Channel**	2	2	2	2	4	4	1	2	4	2
	Wave	2	2	0	1	3	3	1	2	2	1
	Signal	1	2	0	3	2	4	2	3	3	2
	Modulation	0	1	0	2	2	3	0	1	3	1
Ontology2	**Antenna**	1	2	0	3	2	3	1	3	2	1
	Transmitter	0	1	1	2	1	4	0	3	3	1
	Receiver	0	1	0	1	2	2	0	1	3	0
	Filter	2	1	0	2	2	3	1	1	3	1
	Total (STF)	8	12	3	16	18	26	6	16	23	9

After the semantic optimization, the Table III shows the semantic terms for "communication" is - Channel, Wave, Signal, Modulation and for "Hardware" - Antenna, Transmitter, Receiver and Filter.

TABLE IV
Total Term Frequency Distribution

		Wd_it_j		t_js_k		TWd_it_j
|D|	**TF**	**ATF**	**STF**	**ASTF**	**TF+STF**	**ATF+ASTF**
D1	15	0.015789474	8	0.00842105	23	0.02421053
D2	5	0.005263158	12	0.0126315	17	0.01789473
D3	2	0.002105263	3	0.00315789	5	0.00526316
D4	5	0.005263158	16	0.01684211	21	0.02210526
D5	3	0.003157895	18	0.01894737	21	0.02210526
D6	4	0.004210526	26	0.02736842	30	0.03157895
D7	8	0.008421053	6	0.00631579	14	0.01473684
D8	8	0.008421053	16	0.01684211	24	0.02526316
D9	4	0.004210526	23	0.02421053	27	0.02842105
D10	5	0.005263158	9	0.00947368	14	0.01473684

Weight of the document d_i based on only keyword extracted from user query. Where, each Wd_i – is total weight of document d_i and twd_it_j – is total concept in document d_i. The key term and semantic term frequency can be measured as

$$Twd_it_j = \frac{wd_i\ t_j + t_js_k}{dl}$$

The table IV shows the overall weight of each document. t_ja_k – each 't' in document d_i and annotation in each 't_i' except 't'. If keyword 't1' have 3 underlying concepts then a_k=3 and each t_i have a_k annotation references. dl - is total number of words in document d_i.

Where d_i weight of document after adding the semantic terms. From the table I, Query Term Frequency (QTF) shows weight of keyword calculated which is lesser contributed to define right document.

The Semantic Term Distribution (STD) gives real weight of page which contain highest semantic terms related to given query. In a given graph, 'x' axis contains document number and 'y' axis shows overall average frequency terms contains in each document.

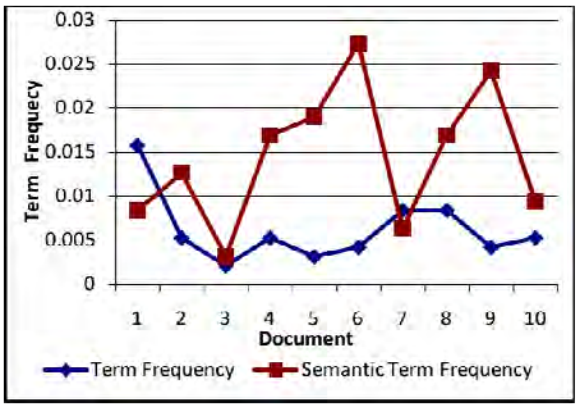

Fig. 2. Semantic term frequency

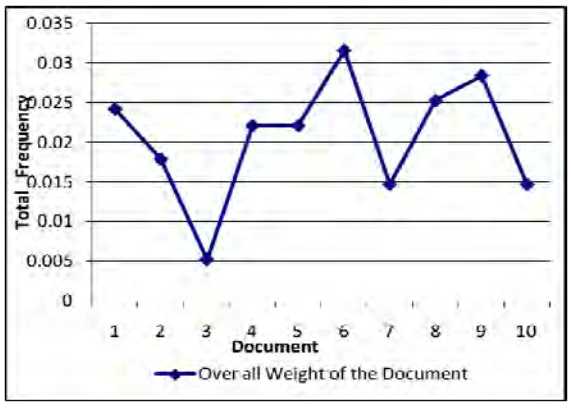

Fig.3. Overall weight of the document

The figure 2 shows the differentiate in weight of the page by means of query and semantic term frequency and the figure 3 shows the overall term frequency distribution in the corresponding document. Form our example documents are d1, d2, d3,…… d10 taken. The semantic term frequency collect document associated to domain ontology's that describe the concepts.

A. Precision

Precision and recall are used to evaluate the performance IR system on document retrieval. Precision 'P' is defined as the proportion of retrieved documents that are relevant, where 'A' is the number relevant document retrieved and 'C' is the number of irrelevant record retrieved. 'B' Relevant record not retrieved and |D| total document retrieved.

- A =7 – Relevant Record retrieved.
- B =1 – Relevant Record Not retrieved.
- C =2 – Number of Irrelevant record retrieved.
- |RD| = 8 – Total number of relevant document.
- |D| = 10 – Total number of records.

To calculate Precision $P = \frac{A \cap |D|}{A+C} * 100$

$$P = \frac{7 \cap 10}{7+2} * 100 = 77.77\%$$

B. Recall

Recall is defined as fraction of the document that is relevant to the user query that is successfully retrieved.

To Calculate Recall $R = \frac{A \cap |D|}{A+B} * 100$

$$R = \frac{7 \cap 10}{7+1} * 100 = 87.5\%$$

The recall shows the better performance of information retrieval system, if the recall increases, then the precision will decrease. Similarly if a precision increase shows that the system's poor performance and recall will automatically decrease considerably.

V. CONCLUSION

In this paper, we proposed ontology as background knowledge for information retrieval. More specifically, this approach addresses two major issue related to ontology; (1) domain ontology selection; (2) filtered the semantically poor terms. The main goal of our approach is to include the value of semantic terms that spread across the document that are directly related with query terms. In our proposed method, identify the user's domain of interest and collect the appropriate domain ontologies based on requirements for extract relevant document. Experimental results shows that the simple method performs better compare to existing information retrieval methods.

REFERENCES

[1] Ofer Egozi, Shaul Markovitch, And Evgeniy Gabrilovich, "Concept Based Information Retrieval Using Explicit Semantic Analysis" ACM Transactions on Information Systems, Vol. 29, No. 2, Article 8, Publication date: April 2011. DOI 10.1145/1961209.1961211 http://doi.acm.org/10.1145/1961209.1961211.A

[2] Danushka Bollegala, Yutaka Matuso, and Mistsuru Ishizuka "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words" IEEE Transaction On Knowledge and Data Engineering, Vol. 23, No. 7, July 2011. DOI: 10.1109/TKDE.2010.172.A

[3] ANK Zaman, Pascal Matskis, Charles Brown "Evaluation of Stop Word Lists in Text Retrieval Using Latent Semantic Indexing", IEEE Sixth International Conference on Digital Information Management. 26-28 September 2011. DOI: 10.1109/ICDIM.2011.6093315A

[4] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, "Annotating Search Results from Web Databases" IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No.3, March 2013. DOI: 10.1109./TKDE.2011.175.A

[5] Hermino Camargo de Souza, Jr., Ana Maria de C, Moura, and Maria Claudia Cavalcanti "Integrating Ontologies Based on P2P Mapping", IEEE Transaction on Systems, Man, and cybernetics – Part A: Systems and humans, VOL. 40.No.5, September 2010. DOI: 10.1109/TSMCA.2010.2044880. A

[6] Feiyu Lin and Kurt Sandkuhl "A Survey of Exploiting WordNet in Ontology Matching" International Federation for Information Processing, Volume 276; Artificial Intelligence and Practice II; Max Bramer; (Boston: Springer), 2008 pp. 341350.A

[7] Jer Lang Hong "Data Extraction for Deep Web Using WordNet" IEEE Transaction on Systems, Man, and Cybernetics – Part C: Applications and Reviews, VOL. 41, No.6, November 2011. DOI: 10.1109/TSMCC.2010.2089678A

[8] Eduard Dragut, Fang Fang, Prasad Sistla Clement Yu, Weiyi Meng "Stop Word and Related Problems in Web Interface Integration" 35th International Conference on VLDB, August 24-28 2009 – Lyon, France.A

[9] Neha Gupta , Saba Hilal "Analysis of Web Content Filtering Factors and the Impact of Sieve Coupons" International Journal of Engineering and Technology, Vol 4 No 4 Aug-Sep 2012.

[10] Iswarya R.J, Bharathi.N "Information Extraction Using Metadata and Solving Polysemy Problems" International Journal of Engineering and Technology, Vol 5 No 2 Apr-May 2013.

[11] Vera Hollink, Theodora Tsikrika, Arjen de Vries "Semantic vs term-based query modification analysis" Proceedings of the tenth Dutch-Belgian Information Retrieval Workshop, Nijmegen, The Netherlands, 2010.

[12] Mannar Mannan j, Sundarambal M and Raguhl "selection of ontology for Web Service Description Language to Ontology Web language conversion", International Journal Computer Science, Volume 10, issue 1 2013. DOI : 10.3844/jcssp.2013.45.53

[13] Senthil Kumar N, Saravanakumar K, "Web Query Expansion and Refinement using Query - Level Clustering" International Journal of Engineering and Technology, Vol 5 No 2 Apr-May 2013.