Web Information Retrieval using WordNet

Jyotsna Gharat
Asst. Professor,
Xavier Institute of Engineering,
Mumbai, India

Jayant Gadge
Asst. Professor,
Thadomal Shahani Engineering College
Mumbai, India

ABSTRACT
Information retrieval (IR) is the area of study concerned with searching documents or information within documents. The user describes information needs with a query which consists of a number of words. Finding weight of a query term is useful to determine the importance of a query. Calculating term importance is fundamental aspect of most information retrieval approaches and it is traditionally determined through Term Frequency -Inverse Document Frequency (IDF).
This paper proposes a new term weighting technique called concept-based term weighting (CBW) to give a weight for each query term to determine its significance by using WordNet Ontology.

General Terms
Term frequency (TF), Inverse Document Frequency (IDF), Vector Space Model, Extraction Algorithm.

Keywords
Information Retrieval (IR), Part of Speech (POS), WordNet, Ontology, Concept-based Term Weighting (CBW).

1. INTRODUCTION
The purpose of information retrieval is to provide information that changes the knowledge state of a user so that this user is better able to perform a present task. An information retrieval process begins when a user enters a query into the system. The information retrieval system compares the query with documents in the collection and returns the documents that are likely to satisfy the user's information requirements. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.
Most IR systems compute a numeric score on how well each object in the database match the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query. Goal of IR is to find documents relevant to an information need from a large document set. Web search engines are the most familiar example of IR systems. Knowledge representation and procedures for processing such knowledge/information [10] are major issues while dealing with information retrieval system.
A fundamental weakness of current information retrieval method is that the vocabulary that searchers use is often not the same as the one by which the information has been indexed. Most of the existing textual information retrieval approaches depend on a lexical match between words in user's requests and words in target objects. WordNet [1, 5, 7 and 8] is a lexical database which is available online and provides a large repository of English lexical items. WordNet is a machine-readable dictionary developed by George A. Miller et al. at Princeton University. In WordNet nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet can also be used for Query Expansion [3].
In proposed method WordNet is utilized to get conceptual information of each word in the given query context. Based on the extracted concepts proposed method can find the weight of a query. Then this is compared with commonly used Vector Space Model using Term-Frequency, Inverse Document Frequency (TF-IDF). The remainder of this paper is organized as follows: Section 2 introduces common approach to find weight of a query. Section 3 discusses proposed method with the help of system architecture. Experiment result is reported in section 4. Finally a conclusion regarding the idea is made in section 5.

2. COMMON APPROACH
Three classic framework models have been used in the process of retrieving information: Boolean, Vector Space and Probabilistic.
Boolean model matches query with precise semantics in the document collection by Boolean operations with operators AND, OR, NOT. It predicts either relevancy or non-relevancy of each document, leading to the disadvantage of retrieving very few or very large documents. The Boolean model is the lightest model having inability of partial matching which leads to poor performance in retrieval of information. Because of its Boolean nature, results may be tides, missing partial matching, while on the contrary, vector space model, considering term-frequency, inverse document frequency measures, achieves utmost relevancy in retrieving documents in information retrieval. The drawback of binary weight assignments in Boolean model is remediated in the Vector Space Model which projects a framework in which partial matching is possible. Vector space model is introduced by G. Salton in late 1960s in which partial matching is possible. TF-IDF [6] is a traditional approach which is used to find the term importance by finding weight of a term.
Steps to find weight of a query using vector space model are as shown in fig 1.

1. Remove punctuation & numbers from web pages.
2. Remove stopwords.
3. Apply Porter stemming algorithm [9].
4. Calculate term frequency (TF) of each term (q) within a query (Q) from document.
5. Calculate Inverse Document Frequency (IDF) of each term in the query (Q).
6. Compute TF-IDF of each term of query using equations (1) and (2).

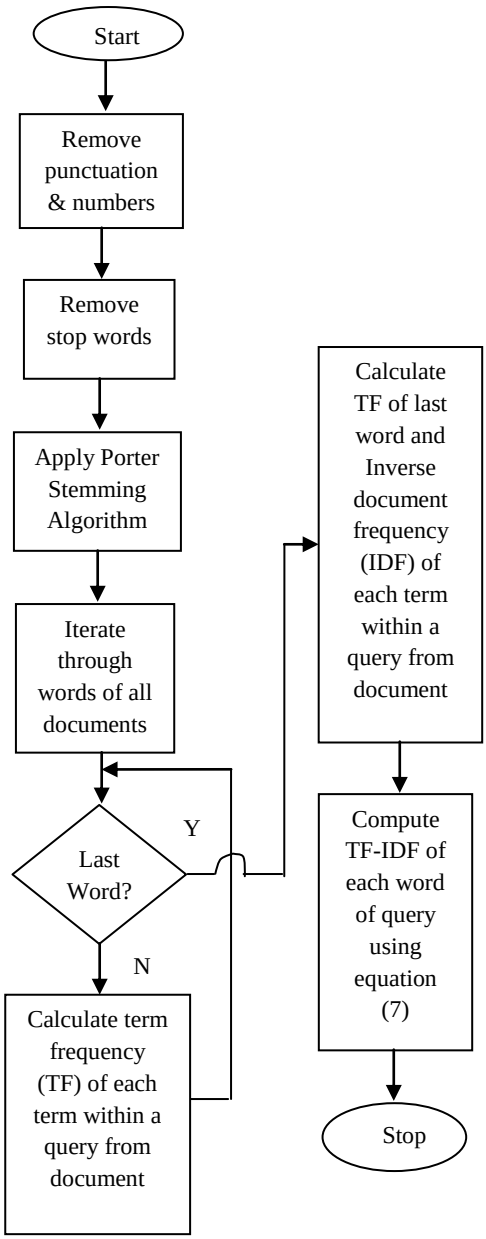

Fig 1: Flowchart for Vector Space Model

Term frequency (TF) is essentially a percentage denoting the number of times a word appears in a document. It is mathematically expressed as shown in equation (1).

$$TF_{q,D} = \left(\frac{\text{Log}(\text{count}_{q,D}+1)}{\text{Log}(\text{numWords}_D+1)}\right) \quad \text{---- (1)}$$

$\text{count}_{q,D}$ = Number of times term q accured in document D
numWords_D = The total number of terms in document D.

Inverse document frequency (IDF) takes into account that many words occur many times in many documents. IDF is mathematically expressed as shown in equation (2).

$$IDF = log\left[N/(n_q + 1)\right] \quad \text{---- (2)}$$

N = Number of documents in the collection
n_q = Number of documents in which term q occurs.

A major drawback of TF-IDF method is that large weighting value may be assigned to rare terms which will lead to invalid classification [4].

3. PROPOSED METHOD

Term significance can be effectively captured using CBW and then be used as a substitute or possible co-contributor to IDF. CBW presents a new way of interpreting ontologies for retrieval, and introduces an additional source of term importance information that can be used for term weighting. In proposed method, Concept-based Term Weighting (CBW) technique is used to calculate term importance by finding the conceptual information of each term using WordNet ontology. The significance of this technique is that

1) it is independent of document collection statistics,
2) it presents a new way of interpreting ontologies for retrieval, and
3) it introduces an additional source of term importance information that can be used for term weighting.

In this research project WordNet is the chosen ontology used by CBW. To determine generality or specificity for a term, conceptual weighting employs four types of conceptual information in WordNet:

1. Number of Senses.
2. Number of Synonyms.
3. Level Number (Hypernyms).
4. Number of Children (Hyponyms/Troponyms).

The term generality vs. specificity can be derived from these 4 types of conceptual information and that term importance can be calculated as a consequence. The more senses, synonyms and children a term has and the shallower the level it appears on, then the more general or vague the term is deemed to be.

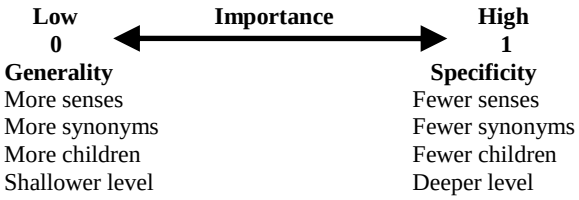

Fig 2: Term Generality vs. Term Specificity

Overview of Concept based term weighting to calculate CBW value of a query term is shown in Fig 3. As shown in figure there are three main steps involved to find the weight of a query. Extraction step extracts conceptual information of each word based on each POS (Noun, Verb, Adjectives) from WordNet. Weighting step find the weight of each extracted integer values for each POS based on weighting functions. After weighting fusion is applied to get a single CBW value for a query term. Any terms used in the query that are non-WordNet terms were given a default high CBW value. This is based on the assumption that the term does not appear in WordNet, is most likely a specific term, and thus it is highly weighted.

The block diagram shown in figure below consists of three main steps:

1. Extraction
2. Weighting
3. Fusion

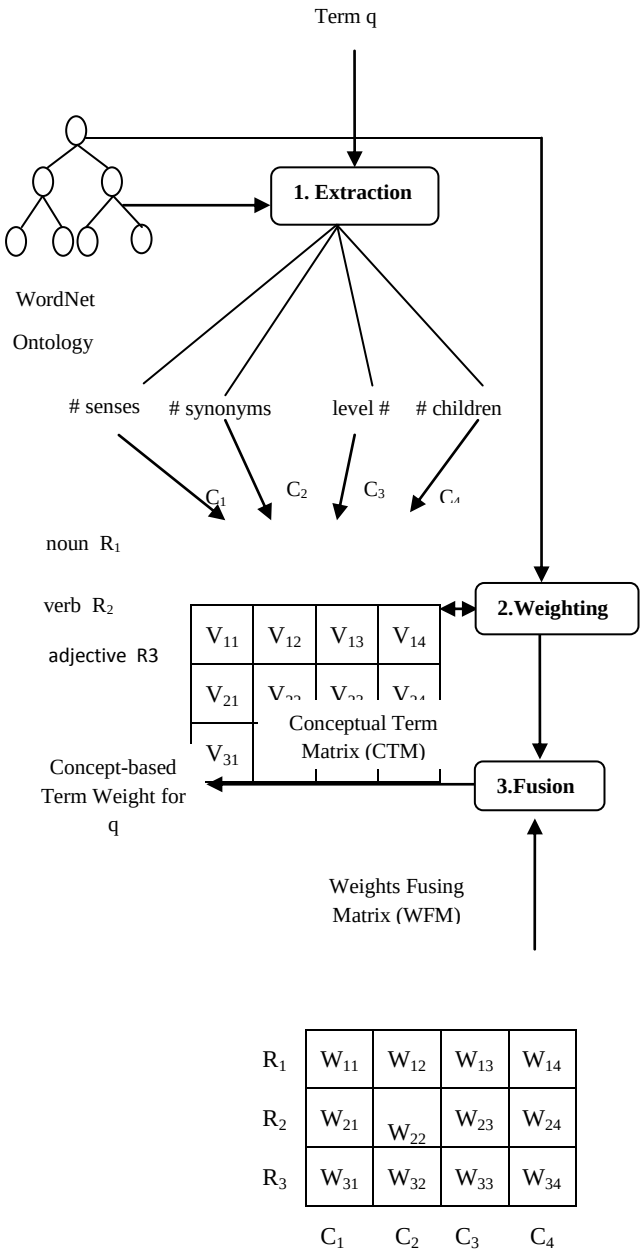

Fig 3: Overview of Concept-Based Term Weighting (CBW)

3.1 Extraction

This step works on a query given by user and extracts the conceptual value for each input query term from WordNet which includes number of senses, number of synonyms, level number (Hypernyms) and number of Children (Hyponyms/Troponyms). Extraction is done by using extraction algorithm [2] as shown below. Initially all values in conceptual term matrix (CTM) are set to -1. Then senses for each POS are counted from WordNet and listed in the first column of CTM. Similarly synonyms for each POS are found by selecting maximum synonyms for senses given by WordNet for a query term. Levels for each POS are found by selecting minimum hypernyms for senses given by WordNet for a input query term and listed in third column of CTM. And finally children for each POS are found by selecting maximum hyponyms/troponyms for senses given by WordNet for a query term. These extracted integer values are stored in Conceptual Term Matrix (CTM).

1. Initialize CTM to (-1).
2. For each row R_m in CTM:
   2.1 Get set of synsets S in R_m section (POS) of WordNet in which q belongs to: S = WordNet (q, POS).
   2.2 Extract conceptual information from S:
   a. V_{m1} = COUNT(S)
   b. V_{m2} = MAX($s_{synonyms}$)
   c. V_{m3} = MIN(s_{level})
   d. V_{m4} = MAX($s_{children}$)

Extraction Algorithm

3.2 Weighting

Weighting is the next step after extraction. Weighting functions convert extracted integer values into weighted values in the range [0, 1]. These weighted values are stored in weighted conceptual term matrix. Based on min, max and avg values for each POS (noun, verb and adjectives) weighting functions are designed as shown in equation (3) and (4). The level number and the number of children are both set to zero for adjectives because adjectives are not organized in a conceptual hierarchy since they are only descriptors of nouns. Therefore, it is not possible to extract the level number and the number of children from WordNet for adjectives. Therefore weighting functions are not created for level number and number of children of adjectives.

a) General Weighting Function for
i. Nouns, Verbs Senses, Synonyms and Children
ii. Adjectives Senses and Synonyms

$$f(x) = \begin{cases} 0 & ,x \geq Max \\ 0.5 & ,x = Avg \\ 1 & ,x = Min \\ f(x-\Delta x) - \dfrac{0.5 * \Delta x}{Avg - Min} & ,Min < x < Avg \\ f(x-\Delta x) - \dfrac{0.5 * \Delta x}{Max - Avg} & ,Max > x > Avg \end{cases}$$

---- (3)

b) General Weighting Function for Nouns, Verbs Levels

$$f(x) = \begin{cases} 0 & ,x = Min \\ 0.5 & ,x = Avg \\ 1 & ,x \geq Max \\ f(x-\Delta x) + \dfrac{0.5 * \Delta x}{Avg - Min} & ,Min < x < Avg \\ f(x-\Delta x) + \dfrac{0.5 * \Delta x}{Max - Avg} & ,Max > x > Avg \end{cases}$$

---- (4)

In above functions Δx is taken as an error factor. These all functions are based on Min, Max and Avg values of each POS. For noun, verb and adjective's senses, weight 0 is assigned for an integer value greater than or equal to Max, weight 0.5 is assigned for an integer value equal to Avg and weight 1 is assigned for an integer value equal to Min. For an integer value in the range [Min, Avg] is given a weight in the range [0.5, 1] and an integer value in the range [Max, Avg] is given a weight in the range [0, 0.5]. Same rules are applied for noun, verb and adjective's synonyms and children. For noun,

verb and adjective's level, weight 0 is assigned for an integer value equal to Min, weight 0.5 is assigned for an integer value equal to Avg and weight 1 is assigned for an integer value greater than or equal to Max. For an integer value in the range [Min, Avg] is given a weight in the range [0, 0.5] and an integer value in the range [Avg, Max] is given a weight in the range [0.5, 1].

3.3 Fusion

Fusion is the last step to get single CBW value of a query that determines the importance of a term. Fusion is performed on weighted conceptual term matrix which is the result obtained by weighting. Fusion considers a new matrix named as Weights Fusing Matrix (WFM) of size 3*4 with all values set to 0.5 to give an average effect. WFM is shown in fig 4.

R_1	0.5	0.5	0.5	0.5
R_2	0.5	0.5	0.5	0.5
R_3	0.5	0.5	0.5	0.5
	C_1	C_2	C_3	C_4

Fig 4: Weight Fusing Matrix (WFM)

Fusing steps:

1. Fuse each column of the weighting CTM with the columns of WFM using column weighted average function.

$$C_n = \frac{\sum_m V_{mn} \times W_{mn}}{\sum W_m}, n = 1,2,3,4 \ldots \quad \text{---- (5)}$$

2. Fuse the row *R* generated in step (1), as shown in fig 5 using row weighted average to give the CBW term importance.

$$CBW_q = \frac{\sum_n C_n \times W_n}{\sum_n W_n} \quad \text{---- (6)}$$

Where *W* is a set of weights with each element being a value in the range [0, 1], and set to 0.5 by default.

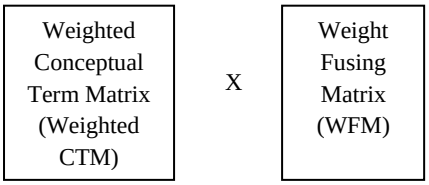

Fig 5: Weighted CTM X WFM

4. EXPERIMENTAL RESULTS

The proposed method is tested by using web dataset which consists of 120 web pages. This dataset is a collection of random web pages. It satisfies the requirement to perform experiment to get term generality or specificity as it provides large collection of real web pages. Preprocessing is performed on web pages to get a clean list of all possible words. Preprocessing involve operations such as removal of all possible stopwords, punctuation and numbers. After that Porter Stemming algorithm is applied on a resultant data. Finally 17209 words are retrieved. These words are used for further analysis.

Using equation (1) and (2) term weight is calculated as shown below:

$$Score_{Q,D} = TF_{q,D} * IDF_q \quad \text{---- (7)}$$

This result gives weight of query using traditional TF-IDF method. Using WordNet CBW value of a query is calculated and final result is listed by using equation (8).

$$Score_{Q,D} = TF_{q,D} * CBW_q \quad \text{---- (8)}$$

Result of equation (7) and (8) is finally compared. Table 1 show the evaluation result where it compares traditional TF-IDF method with proposed TF-CBW method. Total 15 queries Q are fired as input, which consists of 31 query term q and 1 stop word. Stopwords are removed and 31 query terms are used for next processing. 20% queries got TF-IDF value high than TF-CBW. 20% queries resulted in equal values of TF-IDF and TF-CBW. 60% queries resulted in TF-CBW value high than TF-IDF. Based on these results it is clear that proposed method is better than old method.

Table 1. TF-IDF v/s TF-CBW

Sr No	Query	O/P Query (stemmed)	IDF	TF-IDF Score	Avg	CBW	TF-CBW Score	Avg
1	teaching hour	teach hour	0.55 0.33	0.022 0.029	0.026	0.47 0.94	0.019 0.083	0.051
2	a learning course	learn cours	0.90 0.13	0.017 0.022	0.02	0.46 0.75	0.009 0.124	0.066
3	exceptional example	inform commun	1.48 0.93	0.005 0.019	0.012	0.48 0.95	0.002 0.019	0.010
4	program assignment	program assign	0.29 0.33	0.034 0.034	0.034	0.43 0.9	0.05 0.09	0.072
5	tutorial outline	tutori outlin	0.93 1.08	0.022 0.011	0.016	0.72 1.25	0.017 0.013	0.015
6	midterm material	midterm materi	0.78 0.65	0.026 0.023	0.024	0.52 0.9	0.017 0.032	0.024
7	final figure	final figur	0.47 1.60	0.028 0.008	0.018	0.52 0.74	0.031 0.004	0.018
8	neural test result	neural test result	1.38 0.74 1.78	0.009 0.027 0.002	0.013	0.45 0.8 1.2	0.003 0.029 0.001	0.011

9	lecture discussion	lectur discuss	0.44 0.82	0.03 0.018	0.024	0.55 1.01	0.037 0.022	0.030
10	old syllabus	old syllabu	1.60 0.27	0.004 0.024	0.014	0.54 1.09	0.001 0.048	0.024
11	final material	final materi	0.47 0.65	0.028 0.023	0.026	0.52 0.9	0.031 0.032	0.032
12	test design	test design	0.74 0.74	0.027 0.029	0.028	0.35 0.67	0.029 0.026	0.028
13	exam paper	exam paper	0.51 0.90	0.031 0.018	0.024	0.56 0.99	0.034 0.019	0.027
14	office hour	office hour	0.32 0.33	0.031 0.029	0.03	0.36 0.83	0.035 0.083	0.059
15	assign value	assign valu	0.33 1.12	0.034 0.019	0.027	0.37 0.66	0.093 0.011	0.052

Graph as shown below gives the result analysis of Information Retrieval systems with the help of two values, TF-IDF and TF-CBW. Plot of both is shown in fig 6 using query at X-axis and weights at Y-axis.

TFIDF and TFCBW values in the table when plotted in graph show that proposed method is better than the old method.

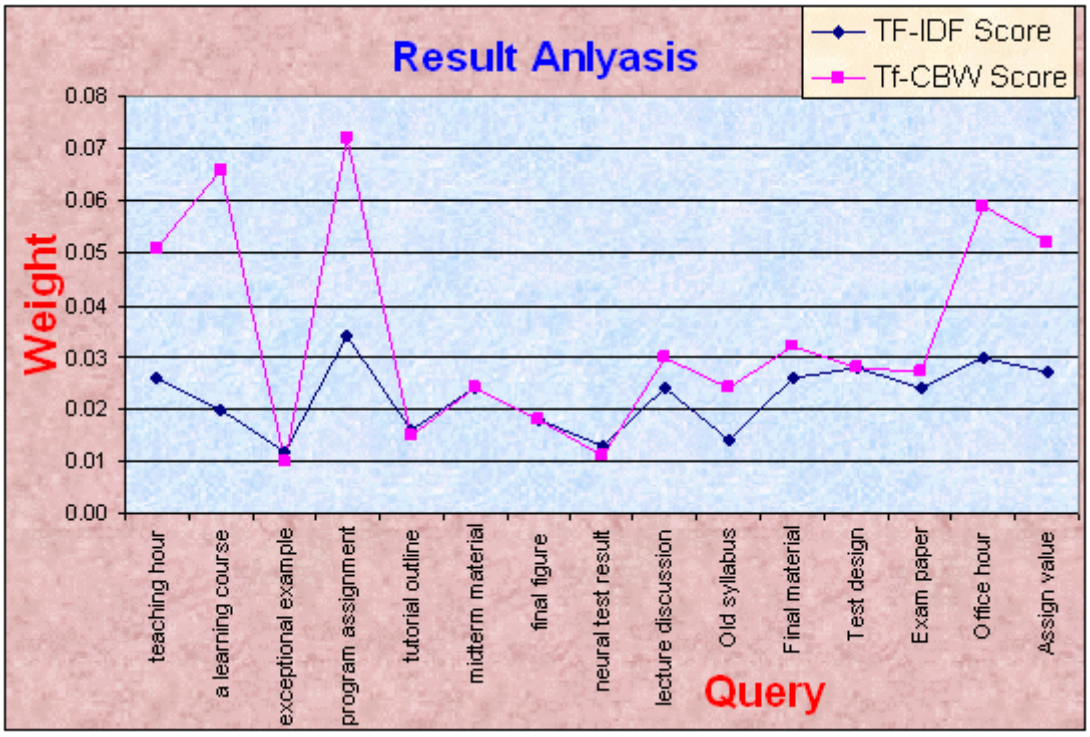

Fig 6: Result Analysis

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

Calculating query term importance was a fundamental issue of the retrieval process. The traditional term weighting scheme TF-IDF approach has following drawbacks:

- Rare terms are no less important than frequent terms – IDF assumption
- Multiple appearances of a term in a document are no less important than single appearance – TF assumption

Because of IDF assumption, the TF-IDF term weighting scheme assigns higher weights to the rare terms frequently. Thus, it will influence the performance of classification. CBW calculates term importance by utilizing conceptual information found in the WordNet ontology. Assumption is made that non-WordNet term should be given high importance of about 0.75 or, generally, in the range [0.5, 1].

As a conclusion, CBW was fundamentally different than IDF in that it was independent of document collection.

The significance of CBW over IDF is that:

1. CBW introduced an additional source of term weighting using the WordNet ontology.
2. CBW was independent of document collection statistics, which is a feature that affects performance.

5.2 Future Work

In the future, above Information Retrieval System can be improved by enhancing the three main components that affect CBW, which are: Extraction, Weighting, and Fusion. Extraction may be enhanced by investigating new types of conceptual information available in the ontology such as: number of attributes, number of parts or causes (Meronyms).

These weighting functions could be investigated to determine another approach for calculating the weighting functions that could potentially lead to better retrieval accuracy. The weights fusing values could be optimized using some other fusion technique. In Future result can be tested and compared by trying different types of ontologies based on conceptual information.

6. REFERENCES

[1] Che-Yu Yang; Shih-Jung Wu, "A WordNet based Information Retrieval on the Semantic Web", Networked Computing and Advanced Information Management (NCM), 2011 7th International Conference, Page(s): 324 – 328, 2011.

[2] Zakos, J.; Verma, B., "Concept-based term weighting for web information retrieval", Computational Intelligence and Multimedia Applications, 2005. Sixth International Conference, Page(s): 173 – 178, 2005.

[3] Jiuling Zhang; Beixing Deng; Xing Li, "Concept Based Query Expansion using WordNet", Advanced Science and Technology, 2009. AST '09. International e-Conference, Page(s): 52 - 55, 2009.

[4] Zhen-Yu Lu; Yong-Min Lin; Shuang Zhao; Jing-Nian Chen; Wei-Dong Zhu, "A Redundancy Based Term Weighting Approach for Text Categorization", Software Engineering, 2009. , Page(s): 36 – 40, 2009.

[5] George A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM, Vol. 38, No. 11, pp. 39-41, 1995.

[6] G. Salton and C. Buckley, "Term – Weighting Approaches in Automatic Text Retrieval", Information Processing and Management, vol. 24, no. 5, pp.513 – 523, 1988.

[7] Measuring Similarity between sentences. [Online]. Available at: http://WordNetdotnet.googlecode.com/svn/trunk/Projects/Thanh/Paper/WordNetDotNet_Semantic_Similarity.pdf.

[8] WordNet Documentation. [Online]. Available at: http://WordNet.princeton.edu/man2.1/wnstats.7WN.

[9] What is Stemming? [Online]. Available at: http://www.comp.lancs.ac.uk/computing/research/stemming/general.

[10] Important problems in information retrieval. Dagobert Soergel, College of Library and Information Services, University of Maryland, College Park, MD 20742, August 1989.