International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 4– No.5,October 2012 – www.ijais.org

Concept based Web Information Retrieval

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

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

ABSTRACT
Information retrieval is concerned with documents relevant to a user's information needs from a collection of documents. The user describes information needs with a query which consists of a number of words. Finding weight of a query is important to determine importance of a query. Calculating term importance is fundamental aspect of most information retrieval approaches and it is commonly determined through Term Frequency- Inverse Document Frequency (TF-IDF).

This paper proposed Concept-based Term Weighting (CBW) technique to determine the term importance by finding the weight of a query. WordNet ontology is used to find the conceptual information of each word in the query.

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
Information Retrieval in most cases is searching relevant information. Searching interesting information is one of the most important tasks in Information Retrieval (IR).
Following issues were identified while dealing with information retrieval system as given in [9].

a. Assisting the user in clarifying and analyzing the problem and determining information needs
b. Knowing how people use and process information
c. Knowledge representation
d. Procedures for processing knowledge/information

An information retrieval process begins when a user enters a query into the system. An IR system accepts a query from a user and responds with a set of documents. Queries are formal statements of information needs, for example search strings in web search engines. The process of retrieving information from the result pages yielded by a search engine is termed as web information extraction.
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. The system returns both relevant and non-relevant material. Generally a search engine presents the retrieved document set as a ranked list of document titles. The documents in the list are ordered by the probability of being relevant to the user's request. The highest ranked document is considered to be the most likely relevant document; the next one is slightly less likely and so on.
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, 6, and 7] is a lexical database which is available online and provides a large repository of English lexical items. Using WordNet ontology the retrieval process can be enhanced by the use of rich vocabulary knowledge in the ontology.
In proposed method WordNet is utilized to get conceptual information of each word in the given query context. Based on the extracted concepts, method can find the weight of a query. Then this is compared with traditional Vector Space Model. The remainder of this paper is organized as follows: Section 2 introduces Vector Space Model (TF-IDF) approach and its drawbacks. Section 3 focuses on WordNet ontology. Section 4 discusses proposed method with the help of system design. Experiment result is reported in section 5. Finally a conclusion regarding the idea is made in section 6.

2. VECTOR SPACE MODEL
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 (TF-IDF) 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 [10] is a traditional approach which is used to find the term importance by finding weight of a term. 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).

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

$count_{\ q,D}$ = 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).

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

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

It is typically combined with term frequency (TF) to form the TF-IDF measure.
A major drawback of TF-IDF method is that large weighting value may be assigned to rare terms which will lead to invalid classification. TF-IDF technique has two assumptions of term frequency distribution in a collection of documents.

1) Multiple appearances of a term in a document are more important than single appearance – *tf* assumption.
2) Terms appearing frequently in many documents have limited-emphasized - *idf* assumption.

The first assumption is acceptable since term frequency (*tf*) take part as a measure to reflect the similarity between samples. As to the second assumption, the inverse document frequency (*idf*) simply takes rare terms as more important than those frequent terms, which is the limitation of *idf* assumption [4].

3. WORDNET

WordNet is a machine-readable dictionary developed by George A. Miller et al. at Princeton University. In lexical based retrieval model many objects relevant to the user query are missed and many unrelated objects are retrieved. Fundamental characteristics of human verbal behavior result in these retrieval difficulties. WordNet ontology can be used to solve this problem. The WordNet ontology is a kind of semantic net that consists of nodes (synsets) that represent unique concepts. These nodes are connected to each other through semantic relations. Such nodes and semantic relations are used for exploring concepts from one to others during the search. Table 1 lists some major semantic relations between concepts defined in WordNet [1, 7].

Table 1. Major semantic relations between concepts defined in WordNet

Semantic Relation	Top	Bottom
Synonymy	*X* is similar to *f(X)*	homo, man, human being, human
Hypernym	*X* is a kind of *f(X)*	Apple *is a kind of* fruit
Hyponym	*f(X)* is a kind of *X*	Zebra *is a kind of* Horse
Holonym	*X* is a part/member of *f(X)*	Wheel *is a part of* a car
Meronym	*X* has part/member *f(X)*	Table *has part* leg
Antonym	*f(X)* is the opposite of *X*	Wet *is the opposite of* dry

WordNet was designed to establish the connections between four types of Parts of Speech (POS) - noun, verb, adjective, and adverb. The smallest unit in a WordNet is synset, which represents a specific meaning of a word. It includes the word, its explanation, and its synonyms. The specific meaning of one word under one type of POS is called a sense. Each sense of a word is in a different synset. Synsets are equivalent to senses which is equal to the structures containing sets of terms with synonymous meanings. Each synset has a gloss that defines the concept it represents [7]. Table 2 represents the current statistics for WordNet 2.1 given by [8].

Table 2. The current statistics for WordNet 2.1

POS	Words	Synsets	Senses
Noun	117097	81426	145104
Verb	11488	13650	24890
Adjective	22141	18877	31302
Adverb	4601	3644	5720
Totals	155327	117597	207016

WordNet and its semantic relatedness measure modules can be used for Query Expansion. In query expansion queries are expanded using well-defined synonymous set in WordNet [3].

4. PROPOSED METHOD

In proposed method, Concept-based Term Weighting (CBW) technique is used to calculate term importance by using WordNet to interpret the conceptual information in ontologies. 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.

WordNet is used to determine the term importance by finding conceptual information for each POS (Noun, Verb, and Adjective). Such conceptual information is used by Concept-based Term Weighting technique to find weight of a query. 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.

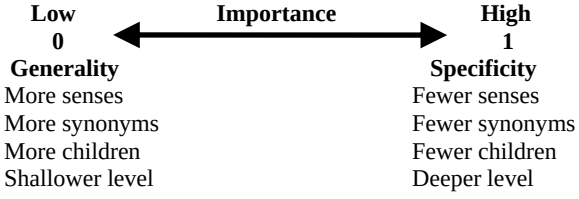

Fig 1: Term Generality vs. Term Specificity

Overview of Concept based term weighting to calculate CBW value of a query term is shown in Fig 2. 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 from WordNet. Weighting step find the weight of each extracted integer values for each POS. These weighted values are stored in Weight Fusing Matrix (WFM). After weighting fusion is applied and final weight of a query term is calculated.

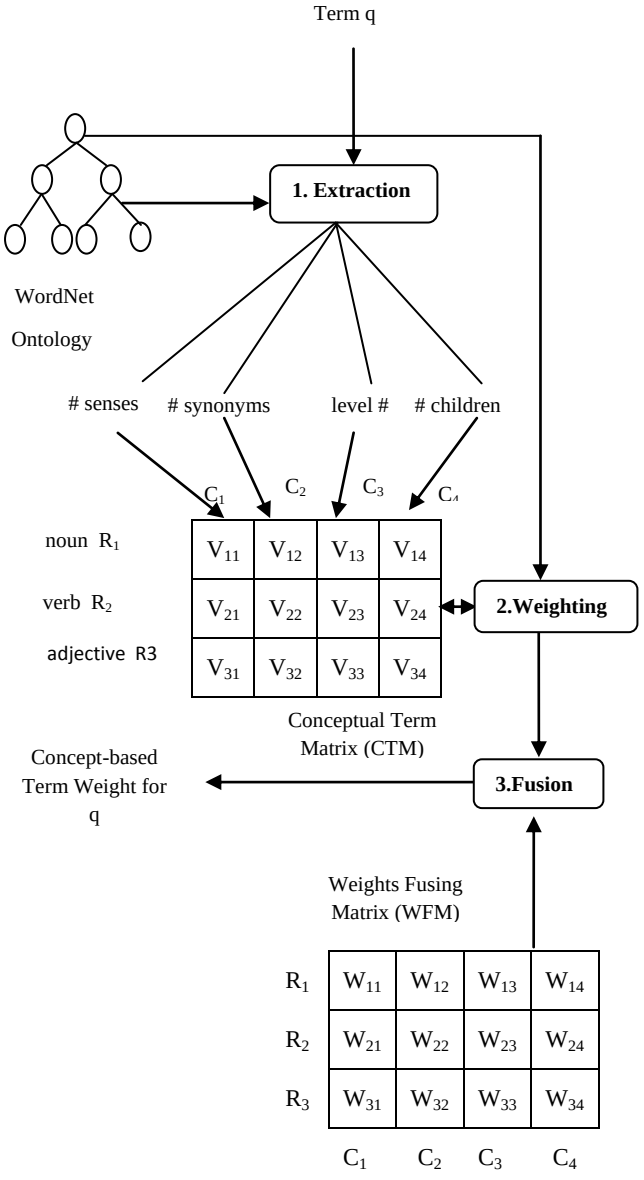

Fig 2: Overview of Concept-Based Term Weighting (CBW)

Above block diagram consists of following three steps:

a. Extraction
b. Weighting
c. Fusion

4.1 Extraction

This step work on the query given by user and extract 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 an 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

4.2 Weighting

Weighting is the next step after extraction. The purpose of a weighting function is to convert each of the extracted integer value into a weighted value in the range [0, 1]. Output of step one that is conceptual term matrix is weighted by using twelve membership function one for each element of the matrix. These functions are based on Min, Max and Avg value of each POS (noun, verb and adjectives) for each type of conceptual information. The general forms of functions are shown as follows:

$$f(x) = \begin{cases} 0 & ,x \geq Max \\ 0.5 & ,x = Avg \\ 1 & ,x = Min \\ f(x-\Delta x) - \frac{0.5 * \Delta x}{Avg - Min} & ,Min < x < Avg \\ f(x-\Delta x) - \frac{0.5 * \Delta x}{Max - Avg} & ,Max > x > Avg \end{cases} \quad \text{-- (3)}$$

General Weighting Function for
a) Nouns, Verbs Senses, Synonyms and Children
b) Adjectives Senses and Synonyms

$$f(x) = \begin{cases} 0 & ,x = Min \\ 0.5 & ,x = Avg \\ 1 & ,x \geq Max \\ f(x-\Delta x) + \frac{0.5 * \Delta x}{Avg - Min} & ,Min < x < Avg \\ f(x-\Delta x) + \frac{0.5 * \Delta x}{Max - Avg} & ,Max > x > Avg \end{cases} \quad \text{-- (4)}$$

General Weighting Function for Nouns, Verbs Levels

In above function Δx is taken as an error factor. Based on Min, Max and Avg values of each POS above functions can be used to find weighted conceptual matrix.

4.3 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. This WFM is fused with weighted CTM. Fusion results in a single and final concept-based term weighting for a query. There are two steps involved in fusion which are as shown below:

Fusing steps:

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

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

5. EXPERIMENTAL RESULTS

The proposed method is tested by using a web dataset which consists of 120 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 [8] is applied on a resultant data. Finally 94537 words are retrieved. These words are used for further analysis.

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

$$Score_{Q,D} = TF_{q,D} * IDF_q \quad \text{-- (5)}$$

This result gives weight of query using traditional TFxIDF method. Using WordNet CBW value of a query is calculated and final result is listed by using equation (6).

$$Score_{Q,D} = TF_{q,D} * CBW_q \quad \text{-- (6)}$$

Result of equation (5) and (6) is finally compared. Table 3 shows the evaluation result where it compares traditional TFxIDF method with proposed TFxCBW method.

Table 3. TFxIDF v/s TFxCBW

Sr No.	Query	O/P Query (stemmed)	IDF	TFxIDF Score	Avg	CBW	TFxCBW Score	Avg
1	functional idea	function idea	0.18 1.04	0.026 0.010	0.018	0.45 0.86	0.066 0.008	0.037
2	a special type	special type	0.50 0.49	0.023 0.023	0.023	0.52 1.03	0.024 0.048	0.036
3	information and communication	inform commun	0.40 0.62	0.029 0.018	0.024	0.42 0.86	0.03 0.025	0.028
4	term weighting technique	Term weight technique	0.32 0.70 0.72	0.025 0.022 0.018	0.022	0.58 1.18 1.7	0.046 0.038 0.043	0.042
5	popular information retrieval	popular inform retriev	1.18 0.40 0.78	0.009 0.029 0.014	0.017	0.45 0.86 1.38	0.003 0.030 0.024	0.019
6	research project	research project	0.38 0.43	0.029 0.023	0.026	0.51 0.86	0.039 0.047	0.043
7	arithmetic theorem	arithmet theorem	0.63 0.50	0.041 0.040	0.040	0.72 1.3	0.046 0.104	0.075
8	contents feature	content feature	0.38 0.66	0.027 0.019	0.023	0.63 1.05	0.045 0.030	0.038
9	suitable example	Suitabl Example	1.0 0.48	0.010 0.030	0.020	0.45 0.93	0.005 0.058	0.032
10	software testing	Softwar Test	0.78 0.72	0.022 0.027	0.024	0.48 1.03	0.013 0.039	0.026
11	set of numbers	Set Number	0.46 0.20	0.031 0.034	0.032	0.2 0.76	0.014 0.131	0.073
12	return average value	Return average valu	0.68 1.48 0.60	0.024 0.004 0.040	0.023	0.34 0.86 1.16	0.012 0.002 0.077	0.030
13	Algebraic language	Algebra Language	0.27 0.54	0.033 0.022	0.028	0.58 0.95	0.072 0.038	0.055

Graph as shown below gives the result analysis of Information Retrieval systems with the help of two values, TFxIDF and

TFxCBW. Plot of both is shown in Fig 3 using query at X-axis and weights at Y-axis.

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

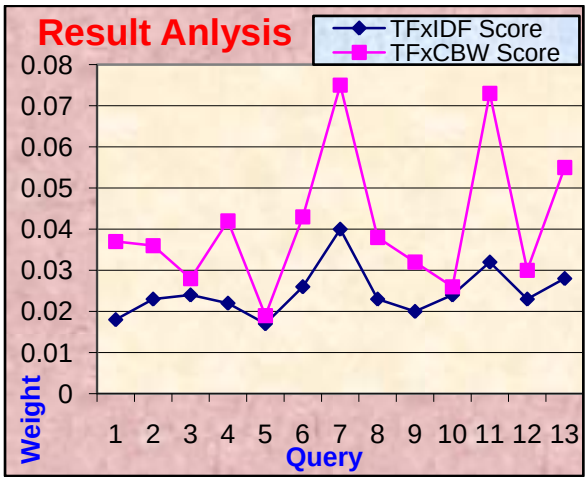

Fig 3: Result Analysis

6. CONCLUSION AND FUTURE WORK

6.1 Conclusion

Calculating query term importance was a fundamental issue of the retrieval process. The TF-IDF term weighting scheme assigns higher weights to the rare terms frequently. Thus, it will influence the performance of classification.

The method is proposed that overcome the above drawbacks of traditional approach. CBW calculates term importance by utilizing conceptual information found in the WordNet ontology. Assumption is made that non-WordNet terms 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.

6.2 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). The 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.

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

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

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

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

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