WEIGHTING QUERY TERMS USING WORDNET ONTOLOGY

Mohammed M. Sakre
Al Shorouk Academy
High Institute for Computers and Information Systems

Mohammed M. Kouta
Arab Academy for Science, Technology & Maritime Transport
College of Computing and Information Technology

Ali M. N. Allam
Arab Academy for Science, Technology & Maritime Transport
College of Computing and Information Technology

ABSTRACT
Term-weighting techniques are responsible for giving a weight for each query term to determine its significance. This term significance is based on the generality or specificity of that term. Traditional weighting models, such as the Inverse Document Frequency (IDF), measure the generality/specificity of terms through document collection statistics. This paper presents a technique that employs the WordNet ontology to determine the significance of query terms without depending on document collection statistics. The experiments are carried out on the WT2g document collection under the Terrier Information Retrieval Platform.

KEYWORDS
Information retrieval, WordNet, ontology, conceptual weighting, term specificity.

1. INTRODUCTION:

The query term weighting technique is responsible for weighting each term in the query submitted to the information retrieval system, indicating the significance of each query term. This is essential so that the ranking models can use this weighting information to calculate the rank scores for the documents. Inverse Document Frequency (IDF) is a statistical scheme, developed by Sparck Jones [8], that determines term specificity according to the number of documents a term appears in relative to the number of documents in the collection:

$$IDF_q = log\left(\frac{N}{n_q}\right) \qquad (1)$$

N: number of documents in the collection
n_q: number of documents where term q occurs.

IDF, and its extensions that depend on the document collection, has become the most popular and important term significance indicator for information retrieval models [1]. Unlike IDF that relies on document collection statistics, concept-based weighting (CBW) that was presented by Verma and Zakos [13] is a weighting technique that interprets conceptual information found in ontologies to determine term specificity or generality.

Manuscript received April 5, 2009
Manuscript revised April 20, 2009

The WordNet® ontology as defined in [15] is a large lexical English database whose structure makes it a useful tool for computational linguistics and natural language processing. WordNet classifies the four parts of speech POS (nouns, verbs, adjectives and adverbs) into synonymous sets (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations such as, Synonyms, Hypernyms, Hyponyms, Troponyms, etc. [2,3,4,5]

2. SYSTEM ARCHITECTURE:

Given a query term q, the system aims to calculate the term importance for q by interpreting the conceptual information related to q in WordNet and giving a resultant value between zero and one. A value close to zero indicates a term with low importance and a value close to one indicates high importance.

The technique is based on the notion that the more general a term is, the less important it should be. Similarly, the more specific a term is, the more important it should be. 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 versus specificity can be derived from these four abovementioned types of conceptual information, and term importance can be calculated consequently. The more senses, synonyms and children a term has, and the shallower the level it appears in, the more general the term is to be considered. Vice versa, the less senses, synonyms, and children a term has, and the deeper the level it appears on, the more specific the term is to be considered. Therefore, as shown in figure (1), general terms are weighted to indicate low importance (i.e. close to zero), and specific terms are weighted to indicate high importance (i.e. close to one).

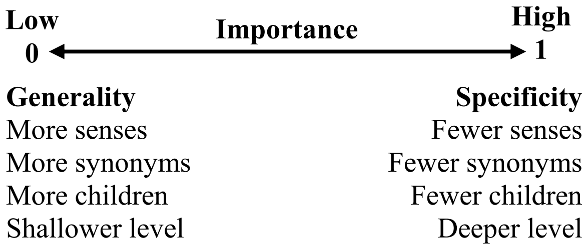

Figure 1: Term Generality vs. Term Specificity

In fact, this is similar to IDF in that the more documents a term appears in, the more general the term is considered to be, and consequently the less important its IDF is calculated.

However, the motivations behind using CBW are based on the shortcomings of IDF or any other term weighting technique that depends on document collection statistics to determine term importance. Moreover, term weighting techniques such as IDF do not always accurately calculate term importance, because it considers a term unimportant just because a term appears in many documents in a collection. Also, with IDF, if a document is added to the collection, it will affect the weight of every single term in the collection because *N* (number of documents in collection) has been increased (equation 1). All these factors make CBW a better alternative to IDF.

As shown in figure (2), CBW performs term weighting as follows. The idea is to first extract this information for a term from WordNet and then proceed to calculate its importance. A (3×4) conceptual term matrix (CTM) is the cornerstone of CBW, in which it holds the four conceptual information types for three different parts of speech represented in the ontology. CBW uses the three parts of speech (POS) sections in WordNet, which are nouns, verbs, and adjectives/adverbs. Both adjectives and adverbs are treated as a single section of POS due to their similar nature and also because the adverbs section has a small size in the WordNet ontology as observed from the WordNet statistics [15].

The conceptual term matrix (CTM) stores confidence indicators for the different conceptual information types that are eventually used to derive a single importance value for a query term.

Each row (R_m) in the CTM represents the different parts of speech, while each column (C_n) represents each conceptual information type. This forms an $(m \times n)$ matrix where m=3 and n=4. Thus, the m^{th} row in the matrix represents a part of speech vector across different types of conceptual information for a single part of speech:

$R_m = \{V_{m1}, V_{m2}, V_{m3}, V_{m4}\}$

On the other hand, the n^{th} column in the matrix is a conceptual information vector across the different parts of speech for a single conceptual information type:

$C_n = \{V_{1n}, V_{2n}, V_{3n}\}$

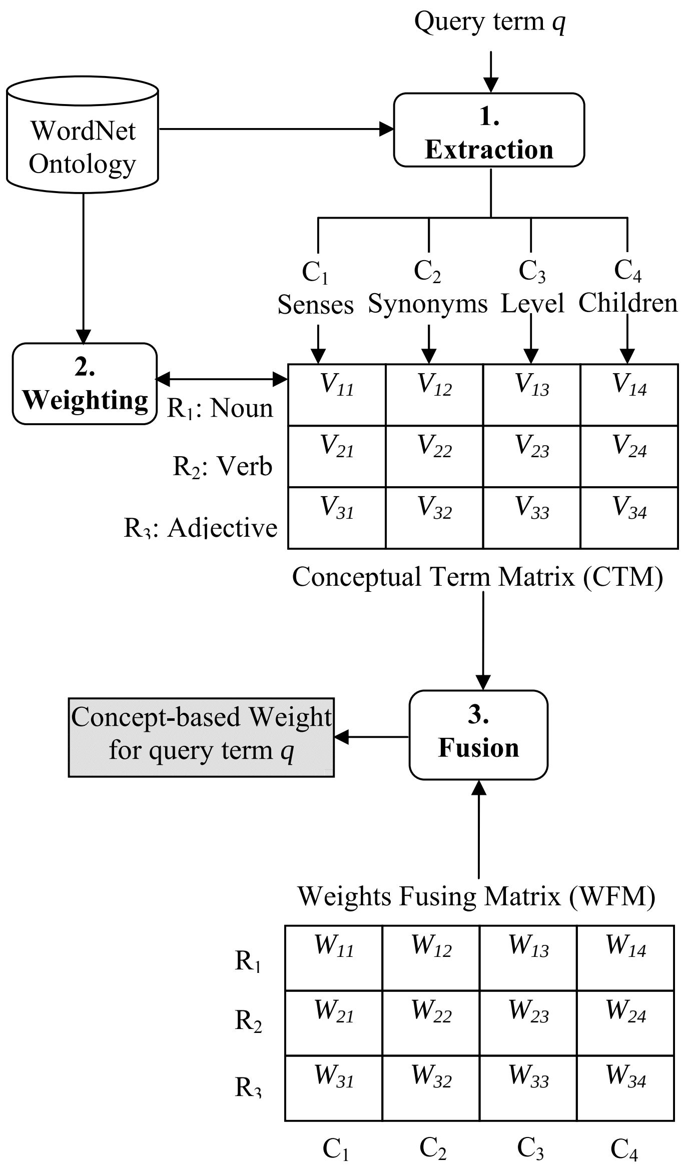

Figure 2: Overview of Concept-Based Term Weighting

Given a query term *q*, CBW can be outlined by the three following steps, and each step will be described in detail throughout the following subsections:

1. ***Extracting*** conceptual information of *q* from WordNet for each POS in the form of integer values and storing them in the CTM.
2. ***Weighting*** the integer values in the CTM to convert them into weighted values in the range [0,1].
3. ***Fusing*** the weighted values in the CTM with WFM to give a single significance weight for the query term *q*.

Any term submitted in the query and not appearing in WordNet is given a default high weight value of 0.75 (i.e. a value close to 1). 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.

2.1. Extraction:

Given a query term *q*, CBW first extracts the four types of conceptual information related to every POS of *q* in the form of integer values, and stores them in the CTM. To perform extraction, term *q* is looked up in the WordNet to determine the set of synsets *S* that *q* belongs to. From these set of synsets, the integer values are obtained and stored for the each of the three other types of conceptual information related to every POS of *q*.

As stated in figure (1), generality is indicated by more synsets, more synonyms, shallower level and more children, and therefore extraction is performed by using the algorithm shown in figure (3):

First, the whole matrix is initialized to (-1). Second, for each POS (i.e. each row), the set of synsets in which *q* belongs to, are determined and located. Using these set of synsets, the integer values are evaluated for the each of the four types of conceptual information related to every POS of *q*.

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

Figure 3: Extraction Algorithm

COUNT(*S*) simply counts the number of synsets in the set of synsets *S* to determine the number of senses. The MAX and MIN functions determine the maximum and minimum values for the three other types of conceptual information (i.e. synonyms, level, and children). After extraction, the resultant CTM, as the example shown in figure (4), will hold integer values for each conceptual information type for every POS for the query term "*waste*":

	C_1	C_2	C_3	C_4
R_1	5	3	4	38
R_2	10	5	1	5
R_3	1	3	0	0

Figure 4: Example of an extracted CTM

The adjective section is very similar to the noun and verb sections, noting that the level number and the number of children are both set to zero. These two values are always set to zero because adjectives are not organized in a conceptual hierarchy since they are only descriptors of nouns. As a result, it is not possible to extract the level number and the number of children from WordNet.

2.2. Weighting:

After the CTM has been extracted, the next step of CBW is to weight the integer values in the CTM. To do this, twelve linear weighting functions are developed, each of which corresponds to an element in the 3×4 matrix.

The purpose of a weighting function is to transform each of the extracted integer values into a weighted valued in the range [0, 1]. This weighted value indicates the term importance where a weight close to (zero) indicates low importance, while a weight close to (one) indicates high importance.

Given the extracted CTM containing integer values, the weighting algorithm is executed as follows:

1. Remove rows in the CTM that are exclusively -1.
2. Convert the remaining integer elements to weights using their corresponding weighting functions.

The most important issue of weighting is determining the weighting functions. There are twelve elements in the CTM, each of which has a corresponding weighting function that takes the integer value stored during extraction and returns its weighted value.

The weighting functions for each of the four conceptual information types can be derived, based on the distribution of terms in WordNet across the three parts of speech, provided by [15] in table (1):

Based on these statistics, ten linear weighting functions were developed to be used as weighting functions for the twelve elements of the CTM, in which the number of children and level number for the adjectives have no weighting functions. After weighting the extracted CTM shown in the previous section (figure 4), the resultant CTM will hold a weighted value in the range [0, 1] for each element, as the example shown in figure (5):

Part of Speech	Conceptual Type	[MIN, MAX]	AVG
Noun	Senses	[1, 7]	2.76
	Synonyms	[0, 7]	1.58
	Levels	[1, 16]	7.5
	Children	[0, 77]	31
Verb	Senses	[1, 7]	3.54
	Synonyms	[0, 7]	1.96
	Levels	[1, 8]	3.64
	Children	[0, 29]	10.8
Adjective / Adverb	Senses	[1, 7]	2.79
	Synonyms	[0, 7]	1.7
	Levels	N/A	N/A
	Children	N/A	N/A

Table 1: MIN, MAX, AVG for POS of WordNet

	C_1	C_2	C_3	C_4
R_1	0.2358	0.3690	0.2308	0.4239
R_2	0	0.1984	0	0.7685
R_3	1	0.3774	0.5	0.5

Figure 5: Example of a weighted CTM

The method of transforming the extracted integer values in the CTM into weighted values works as follows.

For the number of senses, synonyms, and children, an integer value in the range [MIN, AVG] is considered to have a high importance, and therefore given a weight in the range [0.5, 1]. For the level number, an integer value in the range [MIN, AVG] is considered to have a low importance, and therefore given a weight in the range [0, 0.5].

For all the functions, an integer value equal to AVG is given a weight of 0.5. For adjectives, the number of children and level number are always set to 0.5 because adjectives are not organized in a conceptual hierarchy.

The following equations, along with the associated figures (6 to 15), show the linear weighting functions corresponding to the four types of conceptual information (senses, synonyms, levels and children) across the three parts of speech (nouns, verbs and adjectives).

(A) Four Weighting Functions of Nouns:

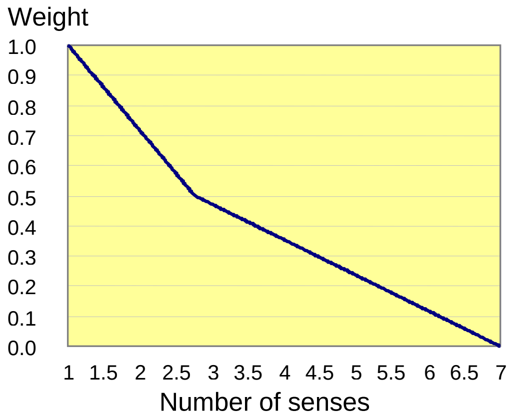

Figure 6:Weighting Function for Nouns Senses

$$
f(x) = \begin{cases} 0 & , x \geq 7 \\ 0.5 & , x = 2.76 \\ 1 & , x = 1 \\ f(x-\Delta x) - \dfrac{0.5}{1.76/\Delta x} & , 1 < x < 2.76 \\ f(x-\Delta x) - \dfrac{0.5}{4.24/\Delta x} & , 7 > x > 2.76 \end{cases}
$$

Equation (2)

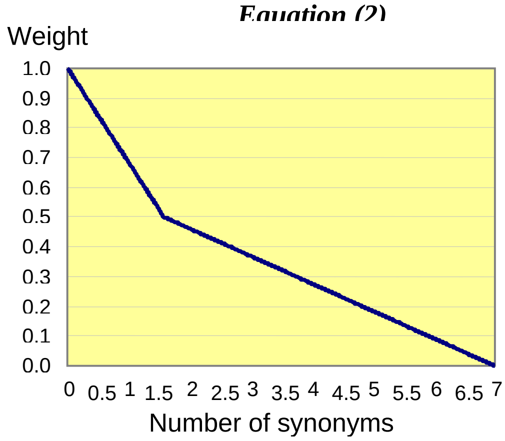

Figure 7: Weighting Function for Nouns Synonyms

$$
f(x) = \begin{cases} 0 & , x \geq 7 \\ 0.5 & , x = 1.58 \\ 1 & , x = 0 \\ f(x-\Delta x) - \dfrac{0.5}{1.58/\Delta x} & , 0 < x < 1.58 \\ f(x-\Delta x) - \dfrac{0.5}{5.42/\Delta x} & , 7 > x > 1.58 \end{cases}
$$

Equation (3)

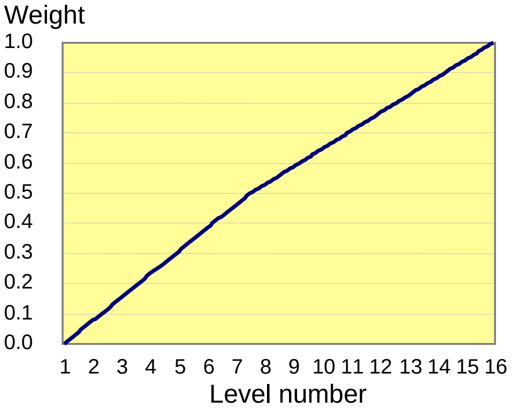

Figure 8: Weighting Function for Nouns Levels

$$
\mathbf{f(x)} = \begin{cases} 0 & , x = 1 \\ 0.5 & , x = 7.5 \\ 1 & , x \geq 16 \\ f(x-\Delta x) + \dfrac{0.5}{6.5/\Delta x} & , 1 < x < 7.5 \\ f(x-\Delta x) + \dfrac{0.5}{8.5/\Delta x} & , 16 > x > 7.5 \end{cases}
$$

Equation (4)

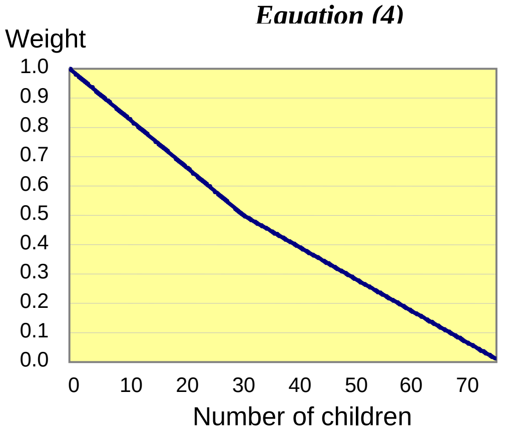

Figure 9: Weighting Function for Nouns Children

$$
\mathbf{f(x)} = \begin{cases} 0 & , x \geq 77 \\ 0.5 & , x = 31 \\ 1 & , x = 0 \\ f(x-\Delta x) - \dfrac{0.5}{31/\Delta x} & , 0 < x < 31 \\ f(x-\Delta x) - \dfrac{0.5}{46/\Delta x} & , 77 > x > 31 \end{cases}
$$

Equation (5)

(B) <u>Four Weighting Functions of Verbs:</u>

Figure 10: Weighting Function for Verbs Senses

$$
\mathbf{f(x)} = \begin{cases} 0 & , x \geq 7 \\ 0.5 & , x = 3.54 \\ 1 & , x = 1 \\ f(x-\Delta x) - \dfrac{0.5}{2.54/\Delta x} & , 1 < x < 3.54 \\ f(x-\Delta x) - \dfrac{0.5}{3.46/\Delta x} & , 7 > x > 3.54 \end{cases}
$$

Equation (6)

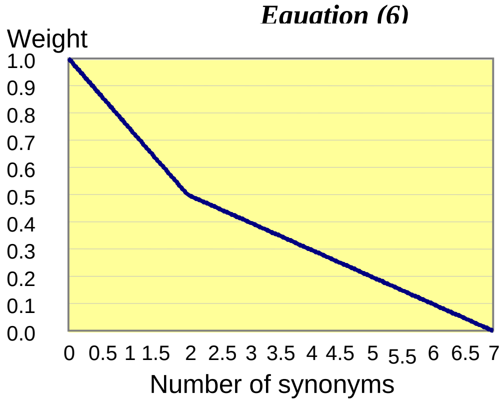

Figure 11: Weighting Function for Verbs Synonyms

$$
\mathbf{f(x)} = \begin{cases} 0 & , x \geq 7 \\ 0.5 & , x = 1.96 \\ 1 & , x = 0 \\ f(x-\Delta x) - \dfrac{0.5}{1.96/\Delta x} & , 0 < x < 1.96 \\ f(x-\Delta x) - \dfrac{0.5}{5.04/\Delta x} & , 7 > x > 1.96 \end{cases}
$$

Equation (7)

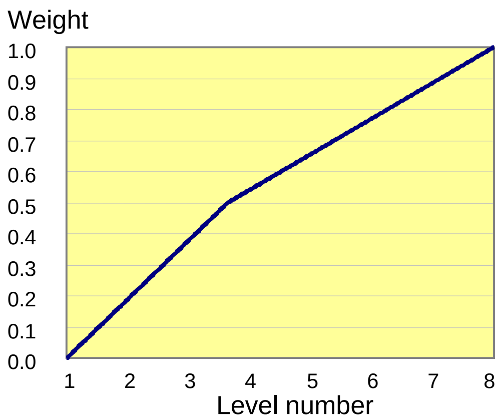

Figure 12: Weighting Function for Verbs Levels

$$
f(x) = \begin{cases} 0 & , x = 1 \\ 0.5 & , x = 3.64 \\ 1 & , x \geq 8 \\ f(x - \Delta x) + \dfrac{0.5}{2.64/\Delta x} & , 1 < x < 3.64 \\ f(x - \Delta x) + \dfrac{0.5}{4.36/\Delta x} & , 8 > x > 3.64 \end{cases}
$$

Equation (8)

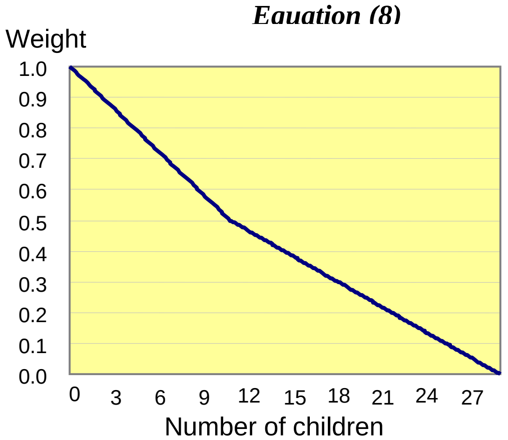

Figure 13: Weighting Function for Verbs Children

$$
f(x) = \begin{cases} 0 & , x \geq 29 \\ 0.5 & , x = 10.8 \\ 1 & , x = 0 \\ f(x - \Delta x) - \dfrac{0.5}{10.8/\Delta x} & , 0 < x < 10.8 \\ f(x - \Delta x) - \dfrac{0.5}{18.2/\Delta x} & , 29 > x > 10.8 \end{cases}
$$

Equation (9)

(C) <u>Two Weighting Functions of Adjectives:</u>

Figure 14: Weighting Function for Adjectives Senses

$$
f(x) = \begin{cases} 0 & , x \geq 7 \\ 0.5 & , x = 2.79 \\ 1 & , x = 1 \\ f(x - \Delta x) - \dfrac{0.5}{1.79/\Delta x} & , 1 < x < 2.79 \\ f(x - \Delta x) - \dfrac{0.5}{4.21/\Delta x} & , 7 > x > 2.79 \end{cases}
$$

Equation (10)

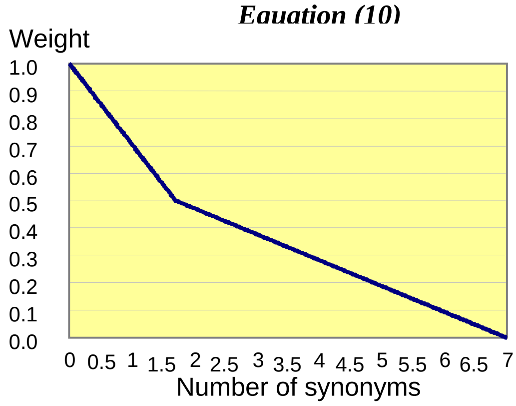

Figure 15: Weighting Function for Adjectives Synonyms

$$
f(x) = \begin{cases} 0 & , x \geq 7 \\ 0.5 & , x = 1.7 \\ 1 & , x = 0 \\ f(x - \Delta x) - \dfrac{0.5}{1.7/\Delta x} & , 0 < x < 1.7 \\ f(x - \Delta x) - \dfrac{0.5}{5.3/\Delta x} & , 7 > x > 1.7 \end{cases}
$$

Equation (11)

2.3. Fusion:

Once weighting has been performed, the CTM is fused as a final step of CBW to give a single term weight that determines the importance of term *q*. Fusion aims to combine all elements in the CTM with the Weights Fusing Matrix (WFM) as shown in figure 16.

	C_1	C_2	C_3	C_4
R_1	0.2358	0.3690	0.2308	0.4239
R_2	0	0.1984	0	0.7685
R_3	1	0.3774	0.5	0.5

Conceptual Term Matrix (CTM)

×

C_1	C_2	C_3	C_4
0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5
0.5	0.5	0.5	0.5

Weights Fusing Matrix (WFM)

Figure 16: Fusing CTM with WFM

All values of the WFM are values in the range [0, 1], and set to 0.5 by default to give an average effect. If there is a row in the CTM that is exclusively (-1), then the corresponding row in the WFM will be set to zero. Fusing is performed using the two following steps:

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

$$C_n = \frac{\sum_m V_{mn} \times W_{mn}}{\sum_m W_{mn}} \text{, at n=1, 2, 3, 4} \qquad (12)$$

This gives a resultant row vector *R*, where each element in the row represents a fused column:
R = {0.4119, 0.3149, 0.2436, 0.5641}.

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

$$CBW_q = \frac{\sum_n C_n \times W_n}{\sum_n W_n} \qquad (13)$$

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. This fusion function has the ability to weight the importance of the different conceptual information types. When used for row fusion, the values of W_1, W_2, W_3, and W_4 correspond to the weights of senses, synonyms, levels, children, respectively. Therefore, the final concept-based weight value for the query term "*waste*" is 0.3837.

3. Experimental Environment:

WT2g [12] was chosen as the benchmark data for testing the system architecture because it fulfills the requirements of an experimental environment, as it provides a relatively large collection of real web pages of 2 GB in size, as well as 50 topics (queries) [10] along with their relevancy judgments [11].

A preliminary experiment was carried out on the WT2g collection by Terrier [6,9] to obtain a result that represents the baseline experiment. Terrier is an open source search engine that is readily deployable on the WT2g collection, and implements indexing as well as retrieval functionalities using the various modern state-of-the-art weighting models such as TFIDF. It also provides different evaluation measures such as: average precision, R-precision, precision at different levels of recall, and number of relevant documents retrieved.

This Terrier baseline run acts as a reference for comparing results that would be achieved by the technique presented in this paper. The evaluation results of Terrier baseline run was obtained using only the original query in the standard TFIDF.

$$score_{Q,D} = \sum_{q \in Q} TF_{q,D} \times IDF_q \qquad (14)$$

where $TF_{q,D}$ is calculated by:

$$TF_{q,D} = \frac{log(N_{q,D} + 1)}{log(DocLength + 1)} \qquad (15)$$

Table (2) shows the evaluation results of Terrier baseline run using only the original query in the standard TFIDF:

Average Precision	Precision at 20	Relevant Documents
0.2900	0.3940	1,871

Table 2: TFIDF Baseline Results

In addition, table (3) along with figure (17) shows the precision figures at different levels of recall for the TFIDF baseline run:

Recall Level	TFIDF Precision
0%	0.761
10%	0.6093
20%	0.4819
30%	0.3938
40%	0.3416
50%	0.3079
60%	0.2291
70%	0.1709
80%	0.1014
90%	0.0399
100%	0.0118

Table 3: Precision at recall levels

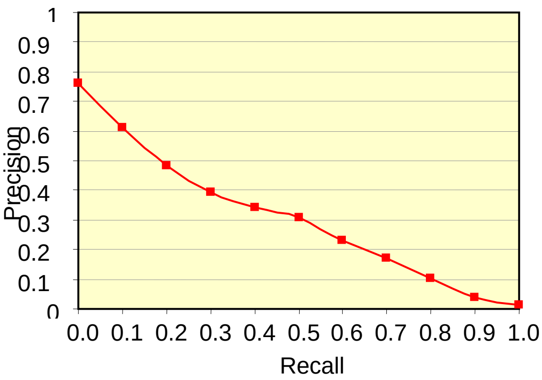

Figure 17: Precision at 11 standard recall levels

3.1. Experimental Setup:

In order to compare CBW against IDF, the rank score of document *D* was calculated using the following retrieval function:

$$score_{Q,D} = \sum_{q \in Q} TF_{q,D} \times CBW_q \qquad (16)$$

where q is a term in the query, and CBW_q is the term importance of q that was replaced instead of IDF_q (equation 14). For all experiments, the query was formed using only the title and description portions of the topic.

3.2. CBW versus IDF:

Both, column fusion and row fusion of the CTM were considered an important step in the process of CBW to give a single term weight value. Column fusion was performed first using the weighted average function (equation 11), in which all weights in the WFM were set to 0.5 giving an average effect for each column. After columns had been fused, the resultant row vector was fused using the weighted average function (equation 13), in which all weights were set to 0.5, resulting in averaging the row vector. Table (4) shows the experimental results of using weighted average functions for row and column fusion.

Average Precision	Δ% (IDF)	Relevant Documents
0.2762	-4.76%	1,681

Table 4: CBW of Term Importance

The CBW run achieved an average precision of 0.2762, which resulted in an accuracy drop of 4.76% from the baseline precision. The individual results of the CBW run, compared to the IDF baseline run, showed that IDF performed more accurately in 27 queries, while CBW performed more accurately in 23 queries. As for the number of relevant documents retrieved, IDF retrieved more documents in 22 queries, and CBW retrieved more documents in eight queries, while the same number of relevant documents was retrieved in 20 queries.

3.3. Non-WordNet Terms:

Another experiment was carried out to estimate the ideal default value for non-WordNet terms. 21 experiments were carried out, testing every value between [0, 1] at intervals of 0.05 for setting the default weight for non-WordNet terms. The results of these experiments supported the assumption that non-WordNet terms should be given high importance of around 0.75, or generally, in the range [0.5, 1]. The best performing run was when the default weight was set to 0.6. This run retrieved 1,690 relevant documents and achieved an average precision of 0.288, which was only 0.69% less than that achieved by IDF baseline.

However, WordNet is not biased toward any specific domain, and this means that it contains only the commonly used terms. Thus, WordNet has a limitation in that its coverage is limited to 155,327 terms [15]. This relatively limited term coverage affected CBW because it considered any term important just because that term did not appear in the WordNet ontology. Therefore, a 0.69% drop in accuracy was insignificant taking in consideration the limited coverage of the WordNet ontology.

3.4. CBW Accuracy:

The previous section showed that CBW achieved the best result using a default value of 0.6 for non-WordNet terms. Figure (18) is an interpolated graph that shows the precision-recall graph of the best achieving CBW against IDF, as observed at table (5).

Recall Level	Precision	
	CBW	Baseline
0%	0.6636	0.761
10%	0.5454	0.6093
20%	0.4454	0.4819
30%	0.3727	0.3938
40%	0.3272	0.3416
50%	0.2772	0.3079
60%	0.2181	0.2291
70%	0.1636	0.1709
80%	0.1272	0.1014
90%	0.1002	0.0399
100%	0.0591	0.0118

Table 5: CBW vs. TFIDF Results

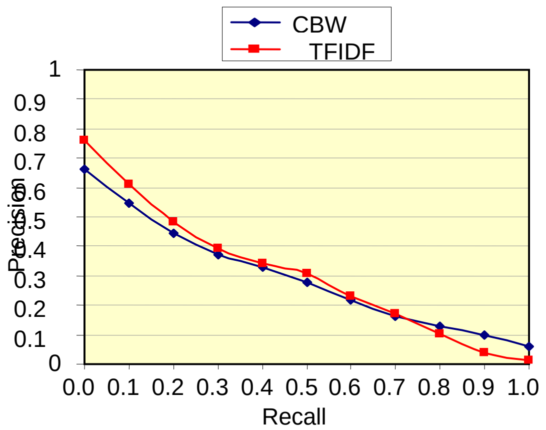

Figure 18: Precision-Recall graph of best CBW vs. TFIDF

It is clear that CBW had a lower precision than the IDF baseline run at low recall levels in the range [0, 0.7]. On the other hand, CBW surpassed the precision achieved by IDF across high levels of recall in the range [0.8, 1], which is in fact an encouraging observation for CBW.

4. Conclusion:

Calculating query term importance was a fundamental issue of the retrieval process. Traditionally, this was determined via Inverse Document Frequency (IDF) by interpreting how many documents a particular term appears in relative to the document collection.

On the other hand, concept-based weighting (CBW) was the technique that has been investigated in this research for calculating term importance by utilizing conceptual information found in the WordNet ontology. These types of conceptual information included the number of senses, number of synonyms, level number, and number of children. The conceptual term matrix (CTM) was the cornerstone in determining a single CBW value for a given term by extracting, weighting, then fusing the conceptual information across all parts of speech (nouns, verbs and adjectives/adverbs).

Finally, the experiment that was carried out to determine the ideal default value for non-WordNet terms supported the hypothesis that non-WordNet terms should be given high importance of about 0.75 or, generally, in the range [0.5, 1], since a value of 0.6 achieved the best results.

As a conclusion, CBW was fundamentally different than IDF in that it was independent of document collection. Experimental results had shown that CBW, compared to IDF, performed with only 0.69% degradation in retrieval accuracy. Although this slight drop in retrieval accuracy, 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. Future Research:

The CBW technique can be optimized by investigating some approaches for enhancing the three main components that affect CBW, which are: Extraction, Weighting, and Fusion.

5.1. Extraction

In this research, CBW extracted four types of conceptual information from the WordNet ontology and stored them in the CTM in the form of integer values. These four extracted conceptual information types were: number of senses, number of synonyms, level number, and number of children.

Therefore, 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), and number of assemblages or entailments (holonyms).

5.2. Weighting

The twelve weighting functions are considered as a significant component of the CBW. These weighting functions were determined through the statistics obtained from the distribution of all WordNet conceptual information types across each POS (Noun, Verb, and Adjective/Adverb). These weighting functions used a straight-forward approach in transforming the extracted CTM integer values into weighted values in the range [0, 1]. These weighting functions could be investigated to determine another approach for calculating the weighting functions that could potentially lead to better retrieval accuracy.

5.3. Fusion

Column and row fusions were carried out using weighted average functions. The weighted average fusing technique was one step in the right direction. However, column weighted average treats all conceptual information evenly, and row weighted average treats all parts of speech evenly. The weights fusing values could be optimized using a technique such as evolutionary algorithm, in order to give higher weights for the more significant elements.

REFERENCES:

[1] Baeza-Yates, R. & Ribeiro-Neto. B., 1999. Modern Information Retrieval. *ACM Press, Addison-Wesley*. p.27-30.

[2] Fellbaum, C., 1990. English Verbs as a Semantic Net. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.278-301.

[3] Gross, D. & Miller, K.J., 1990. Adjectives in WordNet. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.265-277.

[4] Miller, G.A. et al, 1990. Introduction to WordNet: An Online Lexical Database. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.235-244.

[5] Miller, G.A., 1990. Nouns in WordNet: A Lexical Inheritance System. *International Journal of Lexicography, Oxford University Press*. 3 (4), p.245-264.

[6] Ounis, I. et al, 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval.*

[7] Porter, M., 2006. Porter Stemming Algorithm. [Online]. Available at http://tartarus.org/martin/PorterStemmer/ [accessed at 29 February 2008].

[8] Spärck-Jones, K., 1972, 2004. A Statistical Interpretation of Term Specificity and its Application to Retrieval. *Journal of Documentation*, 60 (5), p.493-502.

[9] Terrier 2.1. [Software]. Available at: http://ir.dcs.gla.ac.uk/terrier [Accessed 15 October 2008].

[10] TREC-8 ad hoc and small web topics. [Electronic file]. Available at: http://trec.nist.gov/data/topics_eng/index.html [Accessed 15 October 2008].

[11] TREC-8 small web query relevance judgments. [Electronic file]. Available at: http://trec.nist.gov/data/qrels_eng/index.html [Accessed 15 October 2008].

[12] TREC-8 web collection WT2g. [Electronic web collection]. Available at: http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html [Accessed 15 October 2008].

[13] Verma, B. & Zakos, J., 2005. Concept-based Term Weighting for Web Information Retrieval. *Proceedings of the 6th International Conference on Computational Intelligence and Multimedia Applications (ICCIMA'05), IEEE Computer Society*.

[14] WordNet 2.1. [Software]. Available at: http://wordnet.princeton.edu/obtain [Accessed 26 February 2008].

[15] WordNet Documentation. Available at: http://wordnet.princeton.edu/man2.1/wnstats.7WN [Accessed 26 February 2008].