Send Orders for Reprints to reprints@benthamscience.ae

The Open Automation and Control Systems Journal, 2014, *6*, 473-479

MeSH-based Biomedical Information Semantic Retrieval Model

Qichen Han, Dongmei Li*, Jiaxing Tan, Xuan Wang, Bo Fang and Xuan Tian

School of Information Science and Technology, Beijing Forestry University, Beijing, 100083, China

Abstract: The subject headings is an approach that improves information search accuracy and comprehensiveness to approach multi-language search and intellectualized concept retrieval. Using this method in network information retrieval tool will improve the efficiency of information retrieval. This paper proposes an idea of calculating the similarity based on the relationship among the words in the subject headings. Utilizing query extension, we create a MeSH (Medical Subject Headings)-based Biomedical Information Semantic Retrieval Model (MBISRM). Finally, we compare the results from MBISRM and Baidu in two category realms. The search results from MBISRM are preferable to that of Baidu overall. This paper offers a new stream of thought on applying subject headings in network information system.

Keywords: MeSH, Semantic Retrieval, Similarity Computation, Query Extension, Standardization, Weighted Sort.

1. INTRODUCTION

With the speedy development of the Internet, a significant boom of biomedical information has occurred on the network. It is a big challenge for the industry to conduct information retrieval more effectively and accurately through the incredible amount of data. Currently the major tool for public to use for information retrieval on the Internet is search engine. However, individuals have difficulty to find the most relevant information with such enormous amount of data. The current search engine technology is prepared on the basis of keywords to conduct the search process. A lack of meaningful expression of the words by using this technology makes it problematic to return accurate results that are useful for users.

To solve this problem, some neoteric network information systems and retrieval methods such as conceptual retrieval [1, 2] and semantic retrieval [3, 4] have been proposed. Noumenon is an efficient instrument to realize the semantic retrieval [5, 6], but it takes a lot of work to build and maintain noumenon, and nowadays lots of industrial fields have their own mature subject headings. The subject headings is a relatively developed theoretic system which has been built since 1950s. It becomes an important information organization tool in subject indexing and exerts principal influence in traditional literature indexing and retrieval [7]. Compared with traditional one, our network information retrieval method based on the subject headings focuses more on semantic logic to improve the accuracy. This method has related research in some fields. Using the Metathesaurus designed by The National Library of Medicine of the United States, it enables a syntactic analysis for the input keywords and conducts the query extension through the syntactic analysis result [8]. Peter Clark *et al.* described a knowledge-based Expert Locator application (for identifying human experts relevant to a particular problem or interest), which addresses this issue by using a large technical thesaurus as an initial ontology, combined with simple AI techniques of search, subsumption computation, and language processing [9]. However, neither method mentioned previously has quantitative analysis in relation type of descriptors. For instance, document [10] provides a retrieval method based on agricultural subject headings. However, their method only takes into consideration single stage extension that is directly related to the core search term in querying extension and ignores the influence of other subject words. This article presents a similarity calculation method which draws from the 'Computer similarity by a number of information sources' by Li *et al.* [11], 'Concept vector for similarity measurement' by Hongze Liu *et al.* [12] and incorporates the characteristics of MeSH. This method synthesizes the variety of relations between subject words and designs biomedical information semantic retrieval model (MBISRM) based on MeSH, recurs to the idea of query extension and weighted retrieval. The effectiveness of the model is verified through experiments.

2. MESH-BASED BIOMEDICAL INFORMATION SEMANTIC RETRIEVAL MODEL

2.1. Model Structure

This model contains four modules: subject headings standardization, query extension, webpage crawling and weighted order. First, format the input from users to receive the term K by using the MeSH; second, searching the web information based on the term K; by using the similarity calculation, obtaining the set of terms for query extension and

1874-4443/14

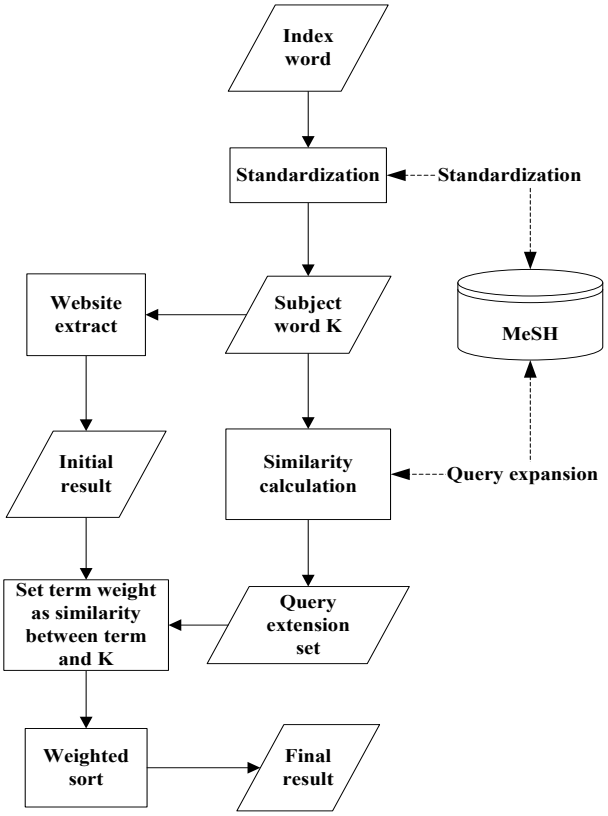

Fig. (1). Model structure.

the corresponding weight. Finally, quantitatively analyzing and sorting grabbed web information based on the terms of query extension and the corresponding weights.

The model structure is shown in Fig. (**1**).

2.2. Standardization

The module extracts user imported keyword then decides whether the term needs to be standardized according to the subject headings. However, users' searching demand and habits are unpredictable. So there are four contingent situations:

①If the term is subject word, there is no standardization required;

②If the term is an entry term in MeSH, it can be transformed into a corresponding subject term through the subject headings;

③If the term matches the subject term partially, the matching terms will be returned for users to select new index term;

④Additional situations will be considered as no query extension required.

2.3. Website Extract

Assuming the subject word K is obtained by standardization, we use the general search engine to conduct the searching process of the term K, then setting the crawl site with the URL of the first s result. Analyzing the URL of s pages through the open source web analytics tool Htmlparse, and extracting the URL, title and content of the web page.

2.4. Query Extension

Using the similarity calculation method to measure similarity of all the terms that is related to K in the subject headings. Selecting the eligible related terms into the query extension set N by setting the threshold value.

2.5. Weighted Sort

When weighted, we set the related terms weight as the similarity between related terms in N and K. The specific algorithm steps of weighted sort are as follows:

Step 1: Calculating the frequency of every related term form the query extension set in the title of the web page(T) and the content(P).

Step 2: Summing up the weights of every web page, and the formula is:

$$TW_n = \sum_{i=1}^{m} W_i \times (\omega \times T_i + P_i) \quad (1)$$

TW_n is the total weight of the nth page; m is the number of related terms in query extension set N; W_i is the similarity between the ith related term in set N and the subject word K; T_i and P_i is the frequency of ith related term in the nth page title and body. ω is the title-text rate, used to adjust the importance of the title in the final result. The larger ω the larger the influence of the title on the page weight is.

Step 3: Sorting the web according to the weight in descending order and returning to the user.

3. SIMILARITY CALCULATION METHOD BASED ON RELATIONSHIP BETWEEN EACH WORD

3.1. Related Definitions

Definition 1 (subject headings concept tree): In subject headings, tree structure T is known as the theme table concept tree, which is constructed by hypernym and hyponym of all keywords which have the top term O, and the root node is O. Node C is named as a subject word node in the tree structure.The number of C's brother nodes is noted as n(C); the depth of root node O is noted as 1; the distance between two nodes is recorded as 1 if the number of branches on the path between them is 1 in the tree.

Definition 2 (shortest path length): The path between two subject word nodes which has the least number of branches is noted as the shortest path length in the T, and the number of branches is the shortest path length.

Definition 3 (closest root node): The subject word node R is the closest root node of A and B, if R is the common ancestor nodes of A,B and the farthest node of the T's root node in all the nodes which conforms to this condition is noted as R (A, B) or R.

Definition 4 (related node): If C belongs to T, and there is a word W that has a correlated relationship with the corresponding subject word with C, then C is the related node of W and W is the related subject word of C.

Definition 5 (ancestor subject word nodes set): The set, which is constituted by all the ancestor nodes of C, is the ancestor subject word nodes set A(C) of C in the T.

Definition 6 (child subject word nodes set): The set, which is constituted by all the child nodes of C, is the child subject word nodes set L(C) of C in the T.

Definition 7 (associated subject word nodes): The set, which is constituted by A(C), L(C) and C itself, is the associated subject word nodes of C in the T.

Definition 8 (node density of associated subject words): In the T, the node density of associated subject words of the root node is 1, and it of the child node is defined as the number of brother nodes +1. Hence, if the node density of associated subject words of C is set to Den(C), then $Den(O)=1$, $Den(C)=n(C)+1$.

Definition 9 (subject word-based density vector): In a T possessing depth of h, the vector $\vec{C}=(V_1,V_2,\ldots V_h)$ is subject word-based density vector. In this vector:

$$V_i=\begin{cases} Den(C_i),\ C_i\subset\{A(C),C\} \\ \delta Den(C_i),C_i\subset\{L(C)\} \\ 0,\quad others \end{cases} \tag{2}$$

C_i is a set of subject word nodes which are the nearest node to C and has the depth of i. δ is the regulatory factor for regulating the influence of density vector to similarity. As the value of δ increases, the influence of the child node to the similarity becomes greater. Smaller the value of δ is, the influence of the brother node to similarity is greater.

3.2. Related Calculation Formula

We stipulate: All the value of the similarity is in [0, 1]. That means if the weight is 0, there is no relationship between the two words; if the weight is 1, two words are equivalent. And if two words required belong to different concept trees, the similarity is 0.

Suppose C1 and C2 need to be determined the similarity, we divide the similarity formula into three categories on the basis of the different relationship types between them: the similarity of equivalence relationship SimD(C1,C2), the similarity of hierarchical relationship SimF(C1,C2) and the similarity of correlated relationship SimW(C1, C2).

(1) Calculating similarity of equivalence relationship

Entry Term and subject word accord with the relation of equivalence----they can be used to replace each other. So

$$SimD(C1,C2)=1 \tag{3}$$

(2) Calculating similarity of hierarchical relationship

$$SimF(C1,C2)=f_1\times f_2\times f_3 \tag{4}$$

f_1 is the similarity based on the shortest path length; f_2 is the similarity based on the depth of the closest root node; f_3 is the similarity based on density. These three similarity calculation methods are as follows:

① similarity calculation based on the shortest path length.

Assuming that in T, the shortest path length between C1 and C2 is d. The similarity calculation formula based on the shortest path length is:

$$f_1(d)=e^{-\alpha d} \tag{5}$$

In formula (5), α is a regulatory factor. f_1 decreases when α is larger.

②similarity calculation based on the depth ofclosest root node.

Setting the depth of R(C1,C2) as h, the similarity calculation formula based on the depth of closest root node is:

$$f_2(h)=1-e^{-\beta h} \tag{6}$$

In the formula (6), β is a regulatory factor. The larger β the larger f_2 is.

③similarity calculation based on the density.

Determining the vector $\overrightarrow{C1}$, $\overrightarrow{C2}$ of C1 and C2 according to the definition 9, the similarity calculation formula based on the density is:

$$f_3(\overrightarrow{C1},\overrightarrow{C2})=\frac{\overrightarrow{C1}\cdot\overrightarrow{C2}}{\left|\overrightarrow{C1}\right|\left|\overrightarrow{C2}\right|} \tag{7}$$

(3) Calculating similarity of correlated relationship

$$SimW(C1,C2)=g_1\times g_2 \tag{8}$$

C1 is the related node of C2. g_1 is the similarity based on the depth of related node. g_2 is the similarity based on the density of related node. The two similarity calculation method as follows:

①similarity calculation based on the depth of related node.

Setting the depth of C1 to h, then we get:

$$g_1(h)=\frac{e^{\varepsilon h}-e^{-\varepsilon h}}{e^{\varepsilon h}+e^{-\varepsilon h}} \tag{9}$$

In the formula (9), ε is a regulatory factor. The larger ε the larger g_1 is.

②similarity calculation based on the density of related node.

Let l be the number of direct child nodes of C1, we get:

$$g_2(b)=1-e^{-\gamma l} \tag{10}$$

In the formula (10), γ is a regulatory factor. The larger γ the larger g_2 is.

3.3. Algorithm for Similarity Calculation

Using the similarity calculation formula given in 3.2, the specific algorithm steps of similarity calculation are as follows:

Step 1: Extending K based on subject headings, we get the initial query extension set of K, denoted by the letter "U", $U=\{D,F,W,Y\}$, in which, D denotes the entry term of K; F denotes all the hypernym/hyponym(all the nodes in the subject headings concept tree T) of K; W denotes the related subject word of K; Y denotes the entry term and related subject word of F.

Step 2: Finding out the top term O and setting it as the root node, then we establish the subject headings concept tree T.

Step 3: Using formula (4) to get the similarity value of F in U and K: $SimF(K,F)$.

Step 4: Using formula (3) to get the similarity value of D in U and K: $SimD(K,D)$.

Step 5 : Using formula (8) to get the similarity value of W in U and K: $SimW(K,W)$.

Step 6: Judging the relationship between every word J in Y and the corresponding subject term I in F: executing step 7 if J is the entry term of I; executing step 8 if J is the related subject word of I.

Step 7: Using formula (3) and (4) to get the similarity between K and J: $SimF(K,I)\times SimD(I,J)$.

Step 8: Using formula (4) and (8) to get the similarity between K and J: $SimF(K,I)\times SimW(I,J)$.

Step 9: Setting the threshold value Q, then deciding whether every word in U has the greater similarity value than Q: if yes adding this word into query extension set N; and skipping this word otherwise.

4. EXPERIMENT AND RESULT ANALYSIS

4.1. Experimental Data

Considering the vocabulary, professional range, experiment requirements, and other factors of the MeSH, we adopt the relationship and subject words in two classes, which comes from the website: http://lib.cqmu.edu.cn/try/cmsh.asp, as the experimental data. They are used to measure the optimal weight of a correlation parameter and evaluate the effect of the relevance ranking.

4.2. The Choice of Index for Evaluating Retrieval Effectiveness

Retrieval effectiveness is defined as: The effective result retrieved by informational retrieval, utilizing retrieval system. It is the measurable indicator of user's satisfaction level with the retrieval result, as well as the direct reflection of retrieval system's performance. In general, the recall ratio and precision ratio are the main evaluation indicator of traditional search engines.

However, some scholars hold another opinion: Around 80% users look only the first page of the result, which means it is more important to the users that the information in demand appears in the first few pages [13-15]. Based on this, some scholars suggest the evaluation indicators on the quality of the sorted search results, and the representative indicator is the search length [16, 17]: the amounts of uncorrelated documents before the nth relative result. This paper selects two indicators to measure the effectiveness of MBISRM retrieval: the correlation of the results and the search length.

Considering that most users check only the first page of the result, we only evaluate the correlation of the first ten results, which are recorded as $P@10$. The calculation method is exhibited by formula (11) and (12).

$$P@10=\frac{a}{a+b} \tag{11}$$

In the formula, a is the number of the results regrading the key words in the top ten search results, and b is the number of irrelevant results. Then we can conclude the formula (12) that the average correlation of the top ten.

$$\overline{P@10}=\frac{\sum_{i=1}^{n}P_i}{n} \tag{12}$$

P_i is the value of $P@10$ in the i time independent experiment.

The search length is defined as the amount of irrelative articles before 5 relative articles, which are recorded as L. Similarly, we can get the average search length formula (13)

$$\overline{L}=\frac{\sum_{i=1}^{n}L_i}{n} \tag{13}$$

L_i is the value of L in the i time independent experiment.

4.3. The Determination of Relevant Parameters Weights

Two important parameters can be determined through experiments. Threshold Q is used for similarity calculation module and the title-text rate ω among weighted sorting module. Other similarity parameters of the algorithm are manually set as: $\alpha=0.2$, $\beta=0.6$, $\delta=0.3$, $\varepsilon=0.6$, $\gamma=0.3$.

In order to make weight measure as accurately as possible, we randomly selected 10 keywords for testing from the experimental data. In the experiment, the website extracting module selects the first 100 search results from Baidu as the general search engine results, and we set the title-text rate to 1 at first. The medical professors confirm the relevance of return result and show the final result in the Table **1**.

Table 1. The determination data of threshold value.

Experimental Contents	Weight		Evaluation index	Average value
	Threshold value	Title-text rate		
The determination of threshold value	1	1	$\overline{P@10}$	3.6
			$\overline{L}$	9.7
	0.8	1	$\overline{P@10}$	4.3
			$\overline{L}$	7.7
	0.6	1	$\overline{P@10}$	5.6
			$\overline{L}$	5.7
	0.4	1	$\overline{P@10}$	5.9
			$\overline{L}$	4.9
	0.2	1	$\overline{P@10}$	8.1
			$\overline{L}$	0.5
	0	1	$\overline{P@10}$	7.7
			$\overline{L}$	0.8

Broken line graph in Fig. (**2**) achieve the data visualization.

The Fig. (**2**) shows that when the threshold value Q is 0.2, the $\overline{P@10}$ can reach the highest data value, which means the first 10 results have the highest relevancy, and the search length is the lowest value, which means it is minimum that the amount of irrelative articles before 5 relative articles. Hence, the threshold value is determined to be 0.2.

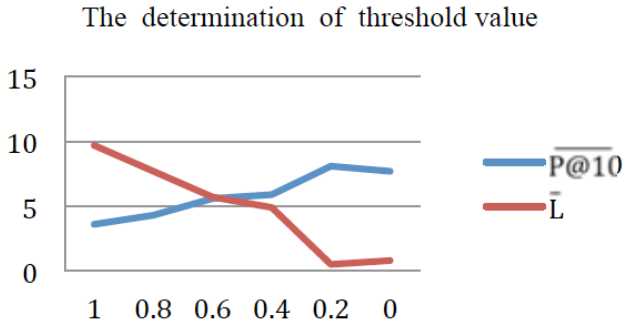

Fig. (2). The determination data of threshold value.

After securing the threshold value(0.2), redoing the test of title-text rate by using the same 10 keywords. The result is showed in Table **2**.

And we draw the broken line graph in Fig. (**3**) and Fig. (**4**) separately.

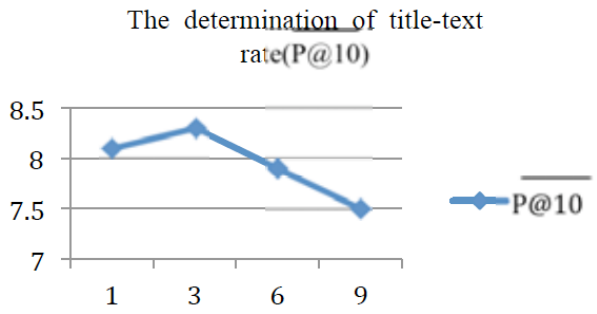

Fig. (3). The determination data of title-text rate ($\overline{P@10}$).

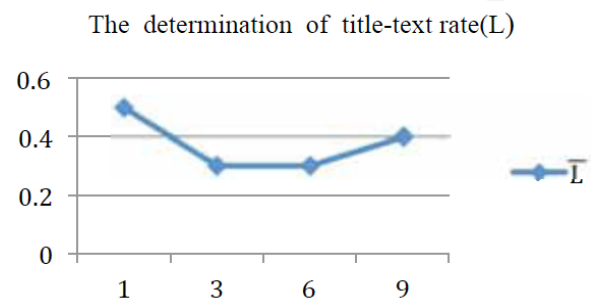

Fig. (4). The determination data of title-text rate($\overline{L}$).

Fig. (**3**) and Fig. (**4**) show: When the value of title-text rate ω is 3, $\overline{P@10}$ can reach the highest value and $\overline{L}$ is one of the smallest. Integrating these two data, we set the title-text rate to 3.

Table 2. The determination data of title-text rate.

Experimental Contents	Weight		Evaluation index	Average value
	Threshold value	Title-text rate		
The determination of title-text rate	0.2	1	$\overline{P@10}$	8.1
			$\overline{L}$	0.5
	0.2	3	$\overline{P@10}$	8.3
			$\overline{L}$	0.3
	0.2	6	$\overline{P@10}$	7.9
			$\overline{L}$	0.3
	0.2	9	$\overline{P@10}$	7.5
			$\overline{L}$	0.4

4.4. Analysis of Experimental Results

According to the optimal weights measured from 4.2, we selected 15 words from the experimental data randomly, retrieving them in Baidu and MBISRM respectively. Then we compare the value of $\overline{P@10}$ and $\overline{L}$; the results are shown in Fig. (**5**) and Fig. (**6**).

According to Fig. (**5**) and Fig. (**6**), it can be observed that the search results from MBISRM is better than that of Baidu overall, which means that subject headings can improve the accuracy of the search results and the retrieval model proposed is feasible in this paper.

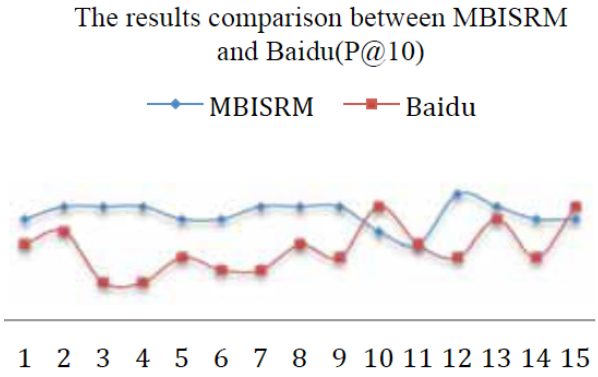

Fig. (5). The results comparison between MBISRM and Baidu(P@10).

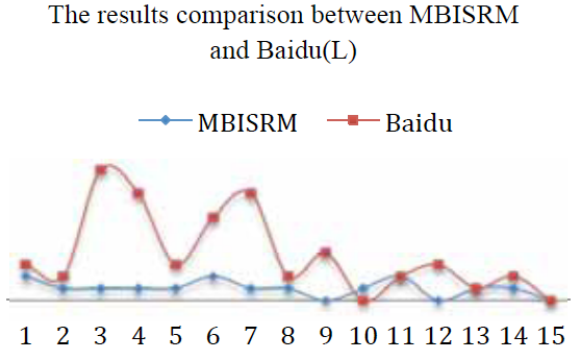

Fig. (6). The results comparison between MBISRM and Baidu(L).

CONCLUSION

The keyword-based traditional information retrieval methods can't adequately express semantic information. Aiming at this kind of defect, we propose a calculation method of semantic similarity between words. Using the relationship between each word in subject headings, we build a biomedical informatics semantic retrieval model based on MeSH and increase the retrieval effectiveness prominently. This model is also suitable for other industries. The retrieval method provides a new research idea on how to use the subject headings reasonably in the era of big data. In future research we can improve and perfect the model from the aspects of retrieval results correlation evaluation and so on.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts.

ACKNOWLEDGEMENTS

This work is supported by the Fundamental Research Funds for the Central Universities (No. xs2014024, No. YX2014-19).

REFERENCES

[1] L. Dianting, and S. M. Ling, "Semantic motion concept retrieval in non-static background utilizing spatial-temporal visual information," *Journal of Semantic Computing*, vol. 7, no. 1, pp. 43-68, 2013.

[2] E. Lotfi, and M.Yaghoobi "Concept retrieval based on a combination of fractal coding, fuzzy rule based system and SVM", *Fractals*, vol. 19, no. 2, pp. 185-194, 2011.

[3] E. M. Van, H. T. V. Schie, and R. A. Zwaan, "The functional role of motor activation in language processing: motor cortical oscillations support lexical-semantic retrieval", *Neuroimage*, vol. 50, no. 2, pp. 665-677, 2010.

[4] P. Daumke, S. Schulz, and M. L. Müller, "Subword-based semantic retrieval of clinical and bibliographic documents" *Methods of Information in Medicine*, vol.49, no. 2, pp. 141-147, 2010.

[5] S. Kara, Ö. Alan, and O. Sabuncu, "An ontology-based retrieval system using semantic indexing", *Information Systems*, vol. 37, no. 4, pp. 294-305, 2012.

[6] A.J.Yepes, R. B. Llavori, and D. R. Schuhmann, "Ontology refinement for improved information retrieval", *Information Processing & Management*, vol. 46, no. 4, pp. 426-435, 2010.

[7] E. Agichtein, and E. Gabrilovich, "Information organization and retrieval with collaboratively generated content", *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 1307-1308, 2011.

[8] A. R. Aronson, T. C. Rindflesch and A. C. Browne, "Exploiting a large thesaurus for information retrieval", *Proceedings of RIAO,* vol. 94, pp. 197-216, 1994.

[9] P. Clark, J .Thompson, and H. Holmback, "Exploiting a thesaurus-based semantic net for knowledge-based search", *Proceedings of AAAI/IAAI*, pp. 988-995, 2000.

[10] X. Xiong, "*Domain Information Retrieval Based on Term Relationships of Thesaurus*", Beijing: Chinese Academy of Agricultural Sciences Dissertation, 2011.

[11] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources", *Knowledge and Data Engineering*, vol. 15, no. 4, pp. 871-882, 2003.

[12] H. Z. Liu, H. Bao, and D. Xu, "Concept vector for similarity measurement based on hierarchical domain sructure", *Computing and Informatics*, vol. 30, no. 5, pp. 881-900, 2012.

[13] B. J. Jansen, "An investigation into the use of simple queries on web IR systems", *Information Research: An Electronic Journal*, vol. 6, no.1, pp. 1-10, 2000.

[14] H. V. Leighton, and J. Srivastava, "First 20 precision among world wide web search services (Search Engines)", *Journal of the American Society for Information Science*, vol. 50, no. 10, pp. 870-881, 1999.

[15] R. Ali, and M. M. Beg, "An overview of web search evaluation methods", *Computers & Electrical Engineering*, vol. 37, no. 6, pp. 835-848, 2011.

[16] M. H. Chignell, J. Gwizdka, and R. C. Bodner, "Discriminating meta-search: A framework for evaluation", *Information Processing and Management,* vol. 35, no. 3, pp. 337-362, 1999.

[17] S. K. Dwivedi, and R. K. Goutam, "Evaluation of search engines using search length", *Proceedings of the International Conference of Computer Modeling and Simulatione*, pp. 502-505, 2011.

Received: September 22, 2014 Revised: November 01, 2014 Accepted: November 06, 2014