# Query Expansion Based on Conceptual Word Cluster Space Graph

Min Peng[1], Quanchen Lin[1] , Ye Tian[1], Ming Yang [1], Yuling Xiao[1], Bin Ni[2]
[1] Wuhan University, Wuhan, China [2] University of South Carolina, Columbia USA
[1] {pengm, quanchen9, tianye, yangming, yuling_xiao }@whu.edu.cn [2]{nib}@engr.sc.edu

***Abstract--*** *For the information retrieval problem, the query expansion methods have been studied for a long time. However, the performance still can't meet the requirements of the users. In this paper, we present a directed graph model, named Conceptual Word Cluster Space Graph (CWCSG), which is constructed to express the semantic similarity among concepts. There are three steps to construct the CWCSG. First, the conceptual words are transformed to different terms. Then, the terms are clustered to different Conceptual Word Clusters (CWCs). After the extraction of CWCs, we calculate the co-occurrence relationship between CWCs to construct CWCSG. Based on CWCSG, user's query is extended to meet user's search requirement in a more accurate way. The experiments show that our method can provide a better performance comparing to classic synonymous dictionary of WordNet.*

## I. INTRODUCTION

With the development of science and technologies, the amount of information on the internet is increasing explosively, which is quite a challenge for the users to locate the useful information. There exists a lot of research works conducted on the information retrieval problem. Generally, most of them are based on the keyword matching. However, most of time it can't satisfy the users' requirement exactly. The shortcomings of keyword matching method are as bellows: first, the search results only include partial information that fulfill user's request; second, the different interpretations of the keyword will return irrelevant information.

Recently, researchers are trying to solve the information retrieval (IR) problem based on natural language processing (NLP), such as POS tagging, syntactic parsing, domain ontology and terminology. Comparing to the keyword matching, the performances of these approaches are improved definitely. But due to limitation of NLP, it still couldn't fully satisfy the users' requirements.

Furthermore, given the fact that the users have various backgrounds, they might not always be able to express their needs accurately for information retrieval. As a result, the expansion of users' query is of great need and particular importance. In this paper, we propose a query expansion method based on a model named *Conceptual Word Cluster Space Graph*, which is short as CWCSG. In *Conceptual Word Cluster Space Graph*, each node refers to a synonym set in a certain document.

The main contribution of this paper is that we propose a novel way to transform a single keyword in the concept space [2] into a set of synonyms with the assistance of WordNet. The proposed scheme improves the efficiency of constructing this special concept space, and also solves the problem of computing similarity redundantly caused by the same meaning words. In this paper, all the synonyms are formed to *Conceptual Word Clusters* as the nodes of CWCSG. The redundant problems are also solved by calculating the co-occurrence relationships among CWCs, which is used to determine the similarity on conceptual level.

The following of the paper is organized as follow. In Section II, the related works are introduced. The *Conceptual Word Cluster Space Graph* is presented in Section III. The experiments and evaluations are conducted and stated in Section IV. Finally, the paper is concluded in Section V.

## II. RELATED WORKS

Query expansion methods have been studied for a long time. Techniques can be divided into global analysis [5][6], local analysis [11], association rules-based [1] [12], user query logs-based [7] and many other query expansion methods [10]. Some methods based on global analysis in query expansion are novel, such as global clustering [5] and

semantic understanding [6]. However, the experiment of global clustering for query expansion is failure. As described by Sparck Jones, they were not able to get anything from clustering after years of hard work [5]. Many methods on query expansion also include the technology of similarity thesauri [7][8] and latent semantic indexing (LSI) [9]. Jing' s Phrasefinder [8] is to build similarity thesaurus. It is a large calculation and also lowers query efficiency because of the words co-occurrence calculation and the pseudo-documents generation. The disadvantage of LSI is also large calculation and it increases recall rate for lowering cost of precision rate. Moreover, LSI can resolve the problem of synonyms, but does nothing with polysemy.

Stefan Riezler proposed a query rewriting method using monolingual statistical machine translation [10]. Monolingual statistical machine translation model highly depends on machine learning materials. Stefan' s method obtains the materials for machine learning from user's log. Similar methods of attaining training data from user's log are quite normal [7][11]. But if the span of the searching content in log files is too big, the training data obtained by the user's log will likely be useless to current query, and it may lead to failure of rewriting. To solve the problems encountered, a graph-based model used for query expansion has been established and achieved a great success. Usually, the position of the two keywords has a great impact on distributed semantic. It means that the semantic similarity from word *A* to word *B* is different from the *B* to *A*'s [3].

## III. CONCEPTUAL WORD CLUSTER SPACE GRAPH

The conceptual graph, which bases on linguistics, psychology, mathematics and philosophy, is used as one of knowledge representation tools. It was introduced in John F. Sowa's Conceptual Structal, which was published in 1984. Afterward, H. Chen, K. J. Lynch proposed a way of regarding concept (keywords or terms) as a node to represent the relationship among concepts [2]. Subsequently, many researchers used the technology for query expansion. A number of major projects (including the University of Illinois Digital Library Project, etc.) used this technology for experiments, designs and had achieved satisfactory results. C. Y. Ng proposed an improved algorithm for constructing concept space in 2001 [3]. However, studies were often based on a single keyword as graph node.

In our CWCSG model, direction edge is used to represent the position relationship between two CWCs. We calculate the similarity of distributed semantic among CWCs based on the *tf***idf* algorithm which was proposed by G. Salton in 1988. Traditional *tf***idf* algorithm calculates on keywords or terms as target, while our method is to figure out CWCs. The CWC can be expressed as a multi-dimensional vector. Each dimension of the vector includes Term and Term frequency. A Term contains both a word that existed in document and the set of the word synonyms in WordNet. Then CWCSG can be constructed by calculating the co-occurrence rate of CWCs.

### *A. Conceptual Word Clusters Extraction*

We pre-process the training documents to pure texts, and tag part of speech (POS). Training dataset is web pages downloaded from the website of http://www.biologynews.net. The words are extracted from POS training dataset to form CWCs. The proposed index algorithm here is different to methods for full-text index [4]. The main steps to build an index in our work are described as follows:

Firstly, extract conceptual words from the document and count their frequency.

Secondly, transform the conceptual word into Term. A Term contains not only the conceptual word itself but also synonyms in WordNet. It is a data structure rather than a string. For example, Terms of conceptual word *predator* and *venoms* are shown in Table I.

Third, once the conceptual word is transformed to Term, Terms-to-concept clustering is followed. Term's clustering is mainly based on the set of synonyms in Term. The same synonym in two Terms means that two Terms are in the same CWC. Each of CWC is given an overall conceptual identification, such as No. 22 shown in Table II, which includes four Terms.

As a result, a document can be expressed by a number of CWCs shown as expression 1.

$D_i$ denotes the $i^{th}$ document, which consists of some CWCs and the sum of the synonyms' frequency for each CWC.

$CWC_i$ is a vector which consists of a number of synonyms, words or terms that express the same meaning (see expression 2). Dimensions in the vector include $Term_i$ and its frequency appeared in the $i^{th}$ document.

*B. CWCSG Construction*

After the extraction of CWCs, the co-occurrence among CWCs is calculated. Each document can be expressed as a matrix. The matrix row contains the CWCs in the document, and the matrix column contains the word or term. All the training dataset is expressed in matrixes. The co-occurrence of CWCs is counted by traversing all the matrixes. The concrete flow chart is shown in Fig. 1.

The associated relationship of CWC *A* with CWC *B* is not equal to the one of *B* with *A*. Therefore, a CWCSG is a directed graph. The weight $W_{ij}$ represents the association value from the end node *i* to the head node *j* (i.e., the similarity from the CWC *i* to *j*), described as (3).

$$W_{ij} = \frac{\sum_{k=1}^{n} D_{kij}}{\sum_{k=1}^{n} D_{ki}} \times weight(j) \qquad (3)$$

$D_{kij}$ in (3) represents the relativity between CWC *i* and *j*, and is described in detail in (4), where $Con_j(n)$ is the number of words or terms in CWC *j*, $tf_{kij}$ is the total frequency of both CWC *i* and *j* appeared in the $k^{th}$ document. $df_{ij}$ is the number of documents both CWC *i* and *j* appeared in the same document.

$$D_{kij} = tf_{kij} \times \log(\frac{N}{df_{ij}} \times Con_j(n)) \qquad (4)$$

$tf_{kij}$ is the smaller frequency between CWC *i* and *j* that appeared in the document, detailed in (5), where $\sum_{i=1}^{n} Coni\_times_i$ is the frequency's summation of all synonymous in CWC *i*.

$$tf_{kij} = \min(\sum_{i=1}^{n} Coni\_times_i, \sum_{i=1}^{n} Conj\_times_i) \qquad (5)$$

$D_{ki}$ in (6) is an extension developed from the classical *tf * idf,* changing the number of the words or terms to CWC as parameter, where $tf_{ki}$ is the number of occurrences of CWC *i* in document *k*, $df_i$ is the number of document in which CWC occurs, $Con_i(n)$ is the number of Terms in CWC *i*, *N* is the number of documents.

$$D_i = \{(CWC_1, count),.....,(CWC_m, count),....,(CWC_n, count)\} \qquad (1)$$

$$CWC_i = \{(Term_1, times),....,(Term_m, times),....(Term_n, times)\} \qquad (2)$$

TABLE I. TERMS OF PREDATOR AND VENOMS

| Term_mainWords | **Predator** |
|---|---|
| Synonyms in WordNet | marauder, predator, vulture, piranha, predatory animal |
| Term_mainWords | **Venoms** |
| Synonyms in WordNet | venom, malice, maliciousness, spite, spitefulness |

TABLE II. NO. 22 CONCEPTUAL WORD CLUSTER

| Concept_ID | 23 | |
|---|---|---|
| Term1 | Term_mainWords | Number |
| | Synonyms in WordNet | number, figure, number, act, routine, turn, bit, number,..... |
| Term2 | Term_mainWords | Numbers |
| | Synonyms in WordNet | Numbers, Book of Numbers, numbers pool, number, ..... |
| Term3 | Term_mainWords | Figures |
| | Synonyms in WordNet | figure, fig, human body, figure, number,..... |

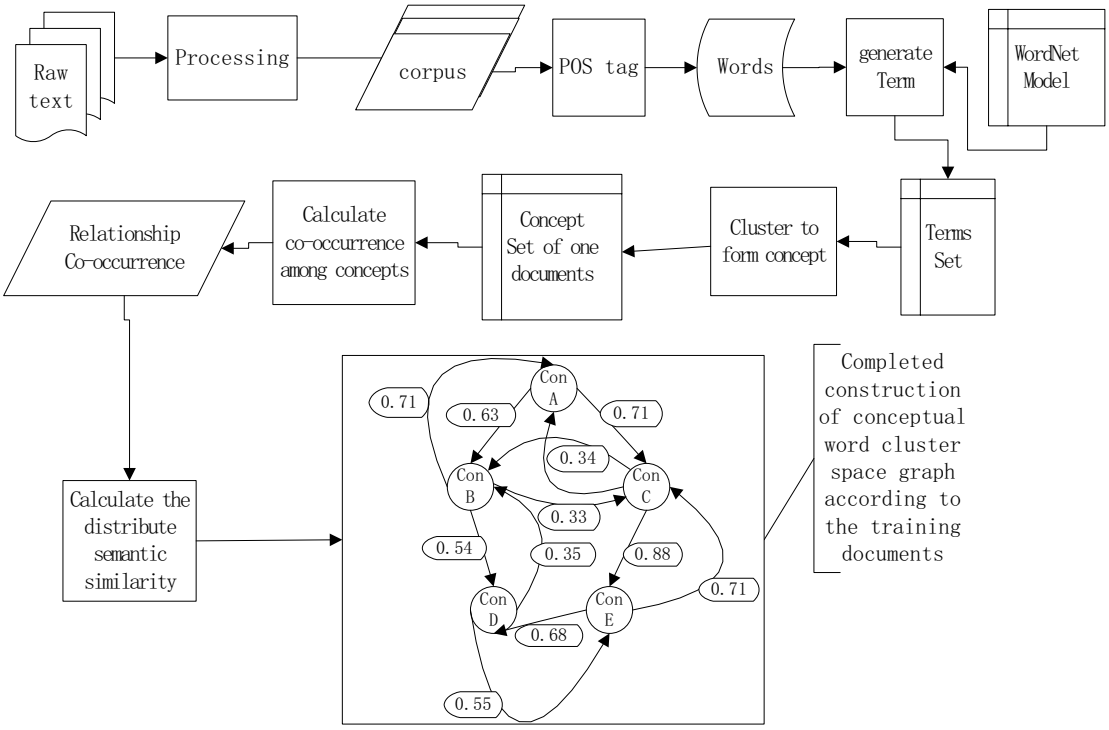


Fig. 1. Concrete Flow Chart of Conceptual Word Cl

uster Space Graph Construction

$$D_{ki} = tf_{ki} \times \log(\frac{N}{df_i} \times Con_i(n)) \qquad (6)$$

*Weight* (*j*) in (3) is a global impact of the $j^{th}$ CWC, described in (7), where $df_j$ is the number of documents which CWC *j* has appeared, *N* is the number of documents.

$$Weight(j) = \frac{\log(N / df_j)}{\log N} \qquad (7)$$

Specific algorithm is described as follows.

Weight *(i; j);*
1. $df_{ij}$ <= 0;
2. $Sum\_tf_{ij}$ <= 0;
3. For each (*k*; $tf_{ki}$) in the ArrayList of *i*
4. **do**
5. **if** there exists (*i*; $tf_{kj}$) in the ArrayList of *j*
6. **then**
7. $df_{ij}$ + = 1;
8. **if** $f_{ki} < f_{kj}$;
9. **then**
10. $Sum\_tf_{ij}$ + = $f_{ki}$;
11. **else**
12. $Sum\_tf_{ij}$ + = $f_{kj}$;
13. $sum\_D_{ij}$ = $Sum\_tf_{ij}$ * log (*N* / $df_{ij}$ + $Con_j$ (n));
14. $sum\_D_i$ = $sum\_tf_i$ * log (*N* / $df_i$ + $Con_i$ (n));
15. *Weight (j)* = log (*N* / $df_j$) / log (*N*);
16. return $sum\_D_{ij}$ * *weight (j)* / $sum\_D_i$;

*C. Query expansion based on the CWCSG*

After CWCSG is established, query expansion's direction is from end node to head node. We define query node as the node that user's query mapped in the CWCSG. Next, extended nodes are selected based on the weights of the edge that links to the query node. User's query is mapped to CWCSG to find the query node. Since association value in CWC *A* with *B* is not equal to CWC *B* with *A*, single-track expanding is more appropriated here. According to the model of CWCSG and the threshold λ specified in advance, the query node is set and treated as the center for node expanding.

Most of the association value among CWCs are zero or very small values in CWCSG. With a threshold λ, our work is to filter the linked CWCs with strong association to query node. Here, strong association is that the association value $W_{ij}$ between two CWCs satisfies $W_{ij}$ >λ. At the same time, not

only the strong association nodes, but also indirect-link nodes which are not linked to the query node directly but also have a strong association to the query node are also added. Whether the indirect-link nodes should be added is determined by indirect-link weight. The indirect-link weight is obtained by multiplying all the reciprocal weights among nodes that help to link the query node to indirect-link node.

CWCs are able to be added into expansion only with strong association value or indirect-link weight. After extension, user's query is extended to many nodes (CWCs). Then, extended query is submitted to the information retrieval engine.

## IV. EXPERIMENT AND EVALUATION

### *A. Corpus Collection*

In our experiments, the training and test documents are both downloaded from the website of http://www.biologynews.net. In our work, we choose all the documents that are released in 2009. It contains 545 documents whose subjects are all about biomedicine, health, and disease documents.

The experiment corpus is divided into two parts. The first part of 413 documents is used for training, and the other 132 documents are used for test. To correctly identify the word or phrase that identifies the concept, we tag POS of training documents with Stanford POS Tagger.

### *B. Performance Evaluation*

According to the 413 training documents, we acquired 3762 CWCs. In the experiment, the calculation for co-occurrence of concepts of 413 training documents took 70 seconds or so, under the Pentium i3 CPU, 4-core processor. The efficiency is acceptable.

Table III showed some distributed semantic value among different CWCs. *i*, *j* is the overall conceptual identification. The cross of *i* and *j* is the distributed semantic value between two CWCs. For example, 0.73188 infers that the 38th CWC associates to the 43$^{th}$ CWC's possibility. The conceptual identification is globally unique and the 38$^{th}$ and 43$^{th}$ CWC is respectively described in Table IV.

Extended query is used for information retrieval in 132 test documents. Conceptual word *X-ray* is the initial query. We expanded *X-ray* under the threshold λ limits to 0.1. The experiments are carried on both CWCSG and WordNet for query expansion respectively. There are 14 documents are related to *radiation*. Retrieval system returns 22 documents where 12 documents are related to *radiation* and other 8 documents are irrelevance under query expansion with CWCSG. While, 18 documents are returned where there are 7 documents are involved with *radiation* under the expansion with WordNet.

We did several queries and acquired performance based on the average results of several experiments. Table VI describes the performance of methods between CWCSG and

TABLE III. DISTRIBUTED SEMANTIC VALUE BETWEEN CONCEEPTUAL WORD CLUSTERS

| i \ j | 43 | 40 | 39 | 33 | 34 |
|---|---|---|---|---|---|
| 38 | 0.731888 | 0.6811423 | 0.6167676 | 0.50266427 | 0.50266427 |
| i \ j | 38 | 40 | 43 | 33 | 34 |
| 39 | 0.16213563 | 0.124805145 | 0.102506764 | 0.07890537 | 0.07697577 |

TABLE IV. 38$^{th}$ CONCEPTUAL WORD CLUSTER

| | | |
|---|---|---|
| Concept_id | 38 | |
| Term | Concept_mainwords | **Shrinkage** |
| | Synonyms in WordNet | Shrinking,shrinkage,shoplifting |
| Concept_id | 43 | |
| Term | Concept_mainwords | Decline |
| | Synonyms in WordNet | Loss, decline, shrinkage, shoplifting, shrinking |

TABLE V. PERFORMANCE OF TWO METHODS

| Methods | Precision | Recall | Fall-Out |
|---|---|---|---|
| CWCSG | 56.55% | 86.41% | 5.78% |
| WordNet | 41.89% | 52.35% | 8.72% |

three points are tested, therefore, precision, recall and fall-out.

Query expansion based on WordNet is just the shallow semantic extension, while CWCSG contains both shallow semantic extension and probability statistics, to establish relationship among CWCs based on CWCs' co-occurrence. Compared to synonymous dictionary of WordNet, the advancement of CWCSG is shown in Table V, which gets quite better results in precision, recall or fall-out.

## V. CONCLUSIONS AND FUTURE WORKS

In this paper, *Conceptual Word Cluster Space Graph* (CWCSG) is proposed to express the semantic similarity among *Conceptual Word Cluster* based on the combination semantic parsing with synonymous dictionary and co-occurrence of concepts. For any information retrieval, the user's query is extended based on the CWCSG to meet more detailed requirement. According to the experiments, the performance of the CWCSG model is much better than that of synonymous dictionary. However, there are also some disadvantages of our method. For example, the construction of CWCSG is based on a synonymous dictionary, which had the problem of ambiguity. If a word has different meanings in different occasions, it is still identified as the same meaning in the construction of CWCSG. This behavior will affect the distributed semantic similarity among CWCs. The future works will focus on how to identify the semantic information of a word in special situation.

## ACKNOLEDGEMENTS

This research has been supported by the *National Science Foundation of China* (NSFC) under Award 61070083. The authors are very grateful for this generous support.

## REFERENCES

[1] Martin Bantista. M. J, Sanchez. D, et al., "Mining Web documents to find additional query terms using fuzzy association rules," *Fuzzy Sets and Systems,* Vol. 148, No. 1, pp. 85-104, 2004.

[2] H. Chen, K. J. Lynch, "Automatic construction of networks of concepts characterizing document databases," *IEEE Transl. J. System, Man and Cybernetics*, Vol. 22, pp. 885-902, Sep 1992.

[3] Chi Yuen Ng, et al, "Efficient Algorithms for Concept Space Construction," in *Proc. 5th Pacific-Asia Conf. Advance in Knowledge Discovery and Data Mining*, Hong Kong, China, 2001, pp. 99-101.

[4] E. M. van Mulligen, M. Diwersy, M. Schmidt, H. Buurman, and B. Mons, "Facilitating networks of information," *in Proc. AMIA Symposium,* Los Angeles, USA, 2000, pp. 868-872.

[5] Sparck. K. Jones, Barber. EO, "What makes an automatic keyword classification effective," *Journal of the American Society for Information Sciences,* Vol. 22, No. 3, pp. 166-175, 1971.

[6] Roberto Navigli,Giuseppe Crisafulli, "Inducing Word Senses to Improve Web Search Result Clustering," *in Proc. 2010 Conf. Empirical Methods in Natural Language Processin,* Massachusetts, USA, 2010, pp. 116-126.

[7] Surdeanu, Ciaramita, Zaragoza, "Learning to rank answers on largeonline QA collections," *in Proc. ACL 2008*, Columbus, USA, 2008, pp. 719-727.

[8] Yufeng Jing,W. Bruce Croft, "An association thesaurus for information retrieval," *in Proc. RIAO 94,* New York, USA, 1994, pp. 146-160.

[9] Dumais. S. T, 1995. "Latent semantic indexing(LSI),TREC-3 report," *in Proc. 3rd Text Retrieval Conf. (TREC-3),* Maryland, USA, 1995, pp.105-115.

[10] Stefan Riezler, Yi Liu, "Query Rewriting Using Monolingual Statistical Machine Translation," *in Proc. ACL 2010,* Uppsala, Sweden, 2010, pp. 569-582.

[11] Kyung Soon Lee, W. Bruce Croft, James Allan, "A Cluster-Based Resampling Method for pseudo-Relevance Feedback," *in Proc. 31th ACM SIGIR conf. Research and Development in Information Retrieval,* Singapore, 2008, pp. 235-242.

[12] Yahia. S. Ben, Jaoua. A, "Discovering knowledge from fuzzy concept lattice," *Studies in Fuzziness and Soft Computing,* Vol. 3, No. 68, pp. 167-190, 2001.