Exploring a Community Clustering Algorithm on Semantic Similarity in Large-Scale Social Network

Laizhong Cui, Yuanyuan Jin, Nan Lu
College of Computer Science and Software Engineering
Shenzhen University
Shenzhen, Guangdong, P. R. China
Email:cuilz@szu.edu.cn, jinyuanyuanzj@qq.com, lunan@szu.edu.cn

***Abstract*-This paper proposes a semantic similarity clustering algorithm on the cluster analysis of large-scale social network. By utilizing the semantic hierarchy of WordNet, the proposed method defines the key concept sets and the concept feature values for the community. In our method, the semantic relations between concepts of the community nodes are also constructed, which expands the application of clustering algorithms from text documents to social network. The cluster structures derived from the proposed algorithm are in concordance with peoples' judgments on a specific area, which will lead to the solution of the clustering problems in the social network of different areas. Compared with VSM and k-MEANS, the experiment results show that the proposed algorithm obtains more reasonable results, which validates its effectiveness.**

Keywords-semantic similarity; WordNet ontology; social network; community structure; clustering algorithm; key concept set

I. INTRODUCTION

Semantic similarity [1][2] is a concept with various definitions according to different areas. Taking the term "virus" as an example, the similarities of virus and its categories differ when it is taken to the biological area and computer area. This is caused by different definitions in these two areas. Therefore, understanding the definition other than the term itself becomes more and more important.

The WordNet ontology [3] is an online word sense mapping system, containing concept word sets from different areas and relationships among them with a sematic network structure. Based on WordNet, this paper extracts words and constructs a semantic IS-A relationship hierarchy and mixes concept word set into the standard hierarchical structure.

It is difficult to calculate similarity about different weighted features and classification learning from the long text and web document processing. The most popular methods for solving the problems are applying to vector space model (VSM) [4][5]. Most of these algorithms are costly in computing power, and some algorithms need some background knowledge, as well as manpower. Therefore, it is difficult to apply these algorithms in the unstructured social networks. Especially, when the nodes with similar meanings have few common words in the community structure, VSM will heavily affect the effect of cluster finding and result in serious deviation from actual structure. For example, the "sports community" and "badminton clubs" should belong to the topic of sports, but the VSM will return 0 in sematic similarity due to no common words.

The social networks are abstractions of complex systems and each node in the social network represents the individual unit in the complex system. The edges between nodes are relationships in the social network formed according to some certain rules. There are various types of social networks in the real world, such as social network, biological network, etc. Finding community structure is not a random selection in a large number of nodes with same properties, but a discrimination in nodes with different types, among which the nodes with same property are linked with more connections, while different types of nodes are sparsely connected. Finding community structures within a social network is an important step towards clustering analysis and research of the network.

Clustering is an important method in finding community structure. Through the clustering method, internal regulations and characteristics can be discovered. The clustering algorithms is capable of automatically generating the category number without adding manual annotation and training classifier. As an unsupervised machine learning

 ISBN: 978-1-61208-419-0

method, clustering has a higher flexibility and better automatic processing power. As the increasing tendency of community information dependence [6][7][8][9], people require intelligent information processing other than the processing of the word pattern or word sense. Therefore, the semantic similarity computing becomes one of the ways to solve community clustering problem. It is crucial in improving the effectiveness and accuracy of the clustering result, judging community structure correlations, classifying communities, and mining data.

This paper proposes a WordNet semantic network learning method. By utilizing the semantic hierarchy of WordNet, the proposed method defines the key concept sets [10] and the concept feature values for the community and then use them to define the semantic similarity. According to the semantic similarity, a community clustering method in social network is presented. The cluster structures derived from the proposed algorithm are in concordance with the judgments of peoples on a specific area, which will lead to the solution of the clustering problems in the social network of different areas. Compared with VSM and k-MEANS [11], the proposed algorithm discovers more reasonable results and shows its effectiveness.

The rest of the paper is organized as follows. Section II provides some relevant knowledge. In Section III, we propose a high performance cluster algorithm (CASN). The experiments and results are presented in Section IV. Finally, Section V concludes our work.

II. RELEVANT KNOWLEDGE

A. WordNet Ontology

WordNet is a widely used English words knowledge base and widely applied in natural language processing, semantic translation, which has attracted much international attention [12]. WordNet is organized by sematic relations. It uses synonym sets (synsets) to representative concepts. Keywords in synsets are bounded and the sematic relationships between synsets are also kept in the hierarchy. One word can be mapped into several synsets and one synset contains several words, which provides a way on representing semantic relationships into the relationships between the concept sets and synsets. WordNet semantic relationships mainly include: the parent and child, synonymous, antonym, is-a-part-of and containment, attribute properties, "leading to" relationships and so on. Based on the English WordNet, the Chinese WordNet is an ontology of the Chinese words and the concept word set, by using existing English-Chinese dictionary library to translate the word in English into Chinese and get the knowledge base. It also has the functions of the concept word, same-word and pan-word. The key concept word is the basic element of Chinese WordNet, and use a number of relation types to connect these concept words, which leads to a key concept word set.

B. Similarity Calculation

The calculation of similarities between any two words starts from mapping the words into the concept word sets that they belong to, and then calculates the similarities between each pair. Finally, according to these similarities of concepts word sets, the word similarities are achieved.

1) The concept word and similarity

For the convenience of knowledge sharing and reuse, WordNet clearly defines concepts of different areas and their relationships. Since there is no formal standard, the descriptions of the same problem in different areas will be different. Even in the same area, different ontologies may also have some heterogeneity. This will significantly affect the utilization of WordNet. The ontology mapping is one of the ways to resolve this heterogeneous problem, while the similarity calculation is a key part of the ontology mapping.

Definition 1 (concept word): Concept is an abstract description of the objects in real world. A concept word is defined as a triad $Con=\{N, A, R\}$, where N is the name, A is the attribute set, R is the relationship set. The name and attribute describe the internal characteristics of the concept, relationship set and express the relation between concept and external environment, and also reflect the external characteristics of concept. Concept words can be represented by instances and therefore be more specific in concept meaning.

Definition 2 (similarity): By the definition 1, it is known that the concept word has three important elements, including name, attribute and relationship. So the similarity

 ISBN: 978-1-61208-419-0

calculation also contains those three components. WordNet is seen as a semantic tree organized according to the hierarchical relationships and concept word relationships. In this paper, the similarity calculation is based on the word content similarity in WordNet, which is the semantics distance between them, in other words it is the path length of two semantics in the semantic tree. Similarity values range from 0 to 1. If two concept words are completely different, the similarity is 0. If they are identical, the similarity is 1. The similarity is calculated by (1), where $len(w_1, w_2)$ is the path length from the word w_1 to the word w_2.

$$SIM(w_1, w_2) = \frac{1}{1 + len(w_1 + w_2)} \quad (1)$$

a) The concept name similarity SIM_{cn}

Usually, the concept word in WordNet is a compound word and it is to determine its semantic distance directly. Therefore, the first several steps will be word segmentation and stopping-word removal to form a stopping-word table and key words extraction. Equation (1) is used to calculate key words similarities and get the summation to represent the similarity of concept name as (2).

$$SIM_{cn} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} SIM(c_{1i}, c_{2j}) \quad (2)$$

where, c_{1i} is the key word of concept c_1, $i \in [1,n]$. c_{2j} is the key word of concept c_2, $j \in [1,m]$.

b)The attribute similarity SIM_a

Attribute consists of attribute names (reflect attribute contents) and attribute domain (attributes' value ranges). Therefore, the attribute similarity calculation must include attribute name similarity and attribute domain similarity, which is described as follows:

$$SIM_a(a_1, a_2) = SIM_{an}(a_1, a_2) + \frac{1}{n} SIM_{ad}(a_1, a_2) \quad (3)$$

$SIM_{an}(a_1,a_2)$ is the attribute name similarity, which can use a similar concept name similarity calculation method to calculate the result. $SIM_{ad}(a_1, a_2)$ is the attribute domain similarity, and it mates calculate.

c) The relationship similarity SIMr

Relationship similarity reflects the connection degree between concept and external. The similarity calculation includes two parts: the relationship name similarity and the relationship association concept similarity, which is described as follows:

$$SIM_r(r_1, r_2) = \frac{1}{n} SIM_{rn}(r_1, r_2) + \frac{1}{n} SIM_{rc}(rc_1, rc_2) \quad (4)$$

$SIM_{rn}(r_1,r_2)$ is the relationship name similarity, achieved by similar calculation method as concept name similarity. $SIM_{rc}(rc_1,rc_2)$ is the association concept similarity, and it is calculated from the concept name similarity of association concept, where rc_1 represents the concept of associate relation r_1, and rc_2 represents the concept of associate relation r_2.

2) The semantic similarity

Semantic similarity calculation principle can be described as follows:

a) According to the relation between parent and child in WordNet, the further the distance between any two concept word nodes is, the smaller the semantic similarity is.

b) The higher density the concept word node locates at, the finer the local concepts are divided, which leads to a lower similarity.

As for two concept word nodes with the same distance, the deeper level it locates at, the more specific it will be, which means that the greater similarity is assigned.

The semantic similarity calculation formula is defined as follow:

$$SIM = \sigma + \alpha \times \bar{\varphi} + \beta \times \bar{\omega} \quad (5)$$

where, α and β are the weights of depth factor and density factor respectively, σ is the distance factor , φ is the density factor and ω is the depth factor.

C. The Key Concept Word Set and Concept Feature Value of Community

In the community clustering process, different community structure or different keyword density in community will lead to the distorted structure. Especially when the community data is in a high dimension, the quality, effect and the calculation speed of the clustering are significantly decreased. In order to improve the efficiency of the clustering, the dimension reduction method is a better choice.

 ISBN: 978-1-61208-419-0

At present, the dimension reduction methods mainly includes TF-IDF, information gain (IG), mutual information (MI), etc. [12][13], which are based on the lexical frequency statistics information. For the convenience of clustering operation, a structuring process for nodes in social network is required, which includes: establishing the community key concept word set and extracting concept feature values, and forming structured documents with key concept words. Similar with the text document processing, the words in structured documents can be divided into two classes: the function words and the content words. The function words are particle, which have no real meaning, while the content words are meaningful. According to the features of content words in the network, such as frequencies, positions and so on, weights are assigned to these words to obtain the concept feature values. This value is propositional to the frequency of the associated concept and if a concept appears in the title of the network, its concept feature value will be increased. When a certain concept feature value is greater than a given threshold, this word can be regarded as a key concept word.

At present, for clustering purpose, a document is always transformed into a noun list and the contribution of words' frequencies to the content of the document is ignored [14], which will lead to an unsatisfying performance. In this paper, the key concept word list illustrated in (6) is utilized, where a social network is regarded as a two-dimension array includes the concept words and their frequencies, to meet the requirement of clustering in the social network.

$$\sigma = \begin{cases} \sqrt{1-\frac{len^2}{\theta^2}}, & len < \theta \\ 0, & len \geq \theta \end{cases} \qquad \varphi = \frac{1}{\ln PN+1}$$

$$\omega = \begin{cases} \frac{\sqrt{|dep - E_d|}}{E_d}, & dep \geq E_d \\ -\frac{\sqrt{|dep - E_d|}}{E_d}, & dep < E_d \end{cases} \qquad (6)$$

In (7), w_i is the i-th concept appearing in community and f_i is the frequency associated with w_i. f_i is calculated by the frequency function, namely:

$$D = \{(w_1, f_1), (w_2, f_2), \ldots, (w_n, f_n)\} \qquad (7)$$

In (8), T_i is the feature value associated with the i-th concept word appearing in community; TF_i is the times which the first i concept word appearing in the community; m_i is the number of communities containing the first i concept word; M is the total number of the concept word in community. From (8), it is obvious that the feature value of a concept word is proportional to the frequency that the concept word appearing in sentence, and in inverse proportional to the number of communities containing the concept word.

$$T_i = TF_i log\left(\frac{M}{m_i}\right) \qquad (8)$$

III. HIGH PERFORMANCE CLUSTER ALGORITHM CASN

At present, no clustering algorithm could be generally applicable to the social network in revealing the complex structures that are represented by all kinds of multi-dimensional data sets. Generally, clustering algorithms can be classified into partitioning clustering, hierarchical clustering and density based clustering. The classic partitioning algorithms are vector space model clustering and k-neighborhood clustering [15], which are efficient for large data sets and applicable to Web document clustering applications. The hierarchical clustering algorithms use the association rules to split or cluster data in a hierarchical form to provide solution for hierarchical clustering. They are mostly applied in small data set.

To address the problems of predefined cluster number, initial value selection and local optimal issue, this paper proposes a clustering algorithm based on semantic similarity, called CASN (Clustering Algorithm of Social Network) to efficiently solve the community clustering problem in social networks.

A. Basic Ideas

The basic idea of our proposed social network community clustering algorithm is to define the node distance between community structures, which represents the similarity between community structures by the node semantic similarity (as shown in Figure 1). According to the similarities, the nodes are clustered one by one, and the closely related clusters will gather into a bigger cluster unit, which will grow in size gradually until all nodes form a cluster.

 ISBN: 978-1-61208-419-0

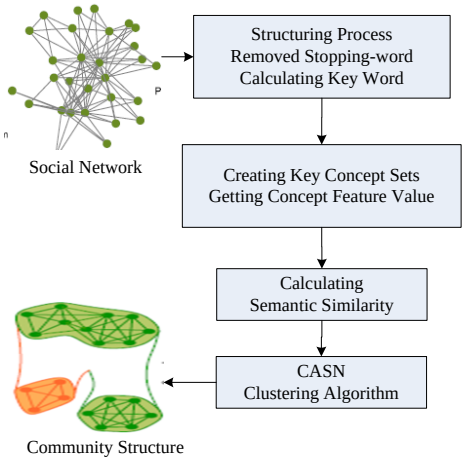

Figure 1. The basic idea of CASN

B. The details of the algorithm CASN

The first step of CASN includes the network structuration and feature extraction. The WordNet concepts and the semantic relationships among concepts are used to generate key concept sets and the concept feature values representing the community structure. Then the clustering algorithm based on the key concept set and the concept feature value is executed. Finally, the key concept set is used to express each clustered unit.

1) The key concept set extraction algorithm CASN-CSET

Clustering Algorithm of Social Network for Concept Set (CASN-CSET) algorithm scans through the identification of each network nodes to extract semantics and maps the extracted nouns into concept by WordNet. Each of the concept is initialized with an interpretation weight. The whole process is described in Figure 2.

Algorithm input: The network node identification document set *D*.

Algorithm output: The key concept of each network node in the document set *D*, *ConSET* [i].

```
i=0; continue=ture;    // i index network node document
/*each network node document in circulation processing D */
do
   file=nextfile (D); //take a network node document in order
   if (isnull(file))
        continue = false;
   else {
      titlewords = gettitle(file);
      word=first(titlewords); //extract the first semantic word
      titlewords.remove(word);
      /* remove the semantic word which have taken out and
        update titlewords */
   }
   while(isnotnull(word)) {
   conceptnode = lookupindexword(word, noun);
   if (isnotnull(conceptnode))
       ConSET[I].add(conceptnode,1);
       /*if concept nodes exist,then join ConSET[i],
         the weight is 1*/
      word = first(titlewords);
      titlewords. remove(word);
   }
   i ++ ;
   while(continue != false)
```

Figure 2. The key concept set extraction algorithm CASN-CSET

2) The concept feature value extraction algorithm CASN-FeaVAl

Clustering Algorithm of Social Network for Feature Value (CASN-FeaVAl) algorithm maps the semantic word into the concept word through the synonym and parent-child relationships in WordNet. Then, a small section of the concept words are selected to represent each document of the network structure. The whole process is described in Figure 3.

Algorithm input: Semantic word concept feature value array obtained from a normalization processing of the network structure document sets, including word segmentation, stemming, stopping and word-frequency calculation.

Algorithm output: Content feature value array *Feat[i]* representing the content of each network structure in the document set *D*.

```
i = 0;   // i index documents
/* circulation handling each network structure document */
do
   for each word in Feat[i] Do
      concept = mapintoconcept(word);
```

```
if (isnotnull(concept)) {
  if (concept in Feat[i])
    add cf to the orginal concepts weight;
    /* cf is concept frequency, the original concept
       feature valueue add concept frequency * /
  else
    Feat[i].add(concept, cf);
    hypernym = getdirecthypernym(concept)
  }
  if (isnotnull(hypernym))
    if (hypernym in Feat[i])
      add cf to the orginal hypernym 's weight;
    /* cf is concept frequency, the original
      concept feature valueue add concept frequency */
    else
   Feat[i].add (hypernym, cf);
   /* put the direct superior concept with concept
      feature valueue */
 endfor
 i ++;
 while (I !=Feat.length)
   i = 0;    /* inetwork node index*/
   /*circulation handleeach network node document */
   do
     for each concept in Feat[i] Do
     hypernym = getdirecthypernym(concept)
    if(isnotnull(hypernym)&&(hypernym in Feat[i]))
       /* reduce dimension */
     if (cf(hypernym) > cf(concept))
       Feat[i].remove(concept, cf)
     else
      Feat[i].remove(hypernym, cf)
     endfor
     i ++;
  while (I != Feat.length)
```

Figure 3. The concept feature value extraction algorithm CASN-FeaVAl

3) The clustering algorithm CASN

Algorithm input: The key concept set *ConSET[i]* and feature value *FeatVAL[i]* of each network structure in document set of network structure, Document number *n*, Cluster number *k*, weight coefficient of *he*, key concept set *kc*, and weight coefficients of concept feature value *cf*. The whole process is described in Figure 4.

Algorithm output: The clustering results *Clusters*, as well as the explanation for cluster results *Results*.

```
P = callsimilarity(ConSET, FeatVAL, kc, cf);
Clusters.initialize (n); .//cluster initialized to n cluster
Results.initialize(ConSET);
/* key concepts set ConSET initialized n cluster explain */
do
  findnearestcluster(c1, c2);
  Clusters. merge(c1, c2); // merger two cluster
  Results.merge(c1, c2);
  /* mark key concept set of explanation cluster and merger,
    and according to the similarity of concept in key concept
    set of cluster results */
  update();  // update similarity matrix
  n --;
  while(n>k) {
    callsimilarity (ConSET, FeatVAL, kc, cf)
    fndnearestcluster(c1, c2)
}
```

Figure 4. The clustering algorithm CASN

The similarity matrix is calculated based on the key concept set and its concept feature vector and its weight coefficient. Then two clusters *c1* and *c2* are found which their similarity degree are the biggest in the clusters.

IV. THE EXPERIMENT AND ANALYSIS

A. Experimental Data

The algorithm's experimental data are taken from 10 discussion groups of a Bulletin Board System (BBS). The discussion groups contain about more than 20000 discussion topics with a total number of 65000 entries.

B. Evaluation Standard

This paper adopts the NMI (Number Mutual Information) analysis method based on mutual information within clusters or categories. As stated by Deng and al. [14], this method can eliminate the influence on the final clustering result caused by the number of clusters. The closer the NMI values are, the better the clustering result is. The NMI value is

 ISBN: 978-1-61208-419-0

calculated as follows:

$$NMI = \frac{\sum_{h,l} n_{h,l} \lg\left(\frac{n n_{h,l}}{n_h n_l}\right)}{\sqrt{\left(\sum_h n_h \lg\left(\frac{n_h}{n}\right)\right)\left(\sum_l n_l \lg\left(\frac{n_l}{n}\right)\right)}} \quad (9)$$

where n_h is the number of data sample in categories h, n_l is the number of data sample in categories l, $n_{h,l}$ is the same data samples in both of the categories h and l, n is the total number of sample data.

C. The Experimental Result

The experiment uses the key concept set and concept feature value to calculate the similarity of the text. *kc* and *cf* are the parameters to adjust the key concept set and concept feature value. Figure 5 shows the effects of clustering by computing NMI under different ratio of *kc:cf*. We can see that when *kc:cf* =1:9, the clustering algorithm has the best performance.

In Figure 6, CASN is compared with other clustering algorithms, set kc:cf =1:9， the NMI value of CASN is about 0.6 when the cluster number is 10; meanwhile, NMI value of VSM algorithm is less than 0.5 but greater than 0.4, and K-means algorithm only get 0.4 on NMI; Even with the cluster number increasing, the NMI values of CASN algorithm are all near 0.6, which is superior to VSM and K-means. The testing results meet the requirements perfectly, which shows the validity and effectivity of the presented algorithms.

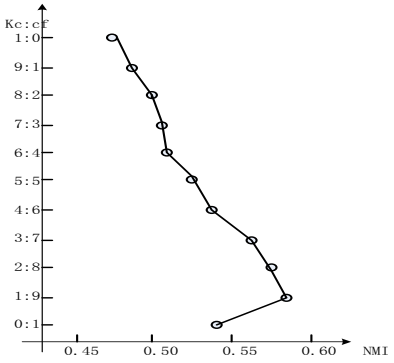

Figure 5. The effects of clustering by computing NMI different ratio of *kc:cf*

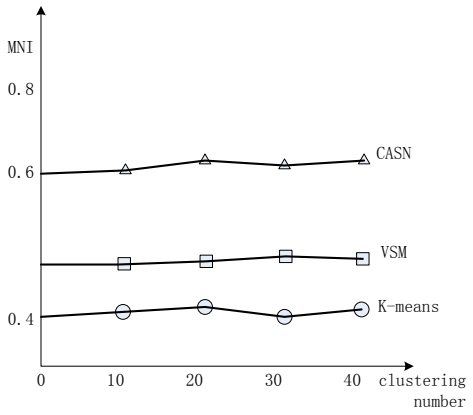

Figure 6. The effects of different clustering methods with *kc:cf* =1:9

V. CONCLUSION AND FUTURE WORK

This paper studied the community clustering algorithm based on semantic similarity under the social network scenario and proposed a feature value using WordNet semantic words which can construct community key concept set to express the community concept. Compared with clustering methods using space vector model SVM, the proposed algorithm shows a better performance. By introducing semantic relations between concept word set (also called synonyms set) and concepts to describe network nodes, our proposed algorithm reduces the dimension of feature lists representing community nodes, and therefore can be applied to the clustering analysis of the community structure. The method proposed in this paper deserves more research on some problems in the future. For example, the hierarchical relationships of the WordNet ontology is still not yet fully utilized. Some fuzzy concept word sets are hard to be defined and a better solution is also required to improve the clustering accuracy.

REFERENCE

[1] R. Zhang, “The Clustering Research On Terminology Definition,” Technology Terminology of China, pp. 14-19, Jan. 2011.

[2] S. Q. Zhao, T. Liu, and S. Li, “A Topical Document Clustering Method,” Journal of Chinese Information Processing, vol. 21, no. 2 pp. 58- 62, 2007.

[3] H. F. Zhu, W. L. Zuo, F. L. He, T. Peng, and W. Y. Ji, “A Novel Text Clustering Method Based on Ontology,” Journal of Jilin University (Science Edition), vol. 48, no.3, pp. 277-285, 2010.

[4] Q. Y. Yao, G. S. Liu, and X. Li, “VSM–based Text Clustering Algorithm” Computer Engineering, vol. 34, no. 9, pp. 39-43, 2008.

 ISBN: 978-1-61208-419-0

[5] M. W. Yuan and P. Jiang, “Compression algorithm for ontology based Vector Space Model Computer Engineering and Applications,” Computer Engineering and Applications, vol. 43, no. 10, pp. 12-17, 2007.

[6] Y. Li, H. Wang, and J. Yang, “Web document clustering algorithm based on semantic similarity,” Journal of hefei university of technology, vol. 32, no. 12, pp. 1846-1850, 2009.

[7] B. L. Yang and K. Y. Shao, “Web document clustering algorithm based on high performance feature selecting function,” Application Research of Computers, vol. 26, no. 2, pp. 546-551, 2009.

[8] P. Zhang, F. Yang, and S. Lu, “Clustering-Based Ontology Block Matching Approach,” Journal of Jilin University (Science Edition), vol. 49, no. 5, pp. 493-499, 2011.

[9] Y. J. Zhang, Y. P. Ren, L. C. Chen, and B. H. Xie, “Component clustering algorithm based on semantic similarity and optimization” Computer Engineering and Design, vol. 31, no.11, pp. 2531-2537, 2010.

[10] J. G. Sun, G. J. Huang, and J. L. LUO, “Modified concept similarity algorithm Computer Engineering and Applications,” Computer Engineering and Applications, vol. 45, no. 5, pp. 154-160, 2009.

[11] T. Kanungo, D. M. MountD, N. S. Netanyahu, C. D. Piatko, R. Siverman, and A. Y. Wu, “A Local Search Approximation Algorithm for K-Means Clustering,” Computational Geometry, vol. 28, no. 2-3, pp. 89-112, 2004.

[12] S. Y. Wu and Y. Y. Wu, “Chinese and English Word Similarity Measure Based on Chinese WordNet,” Journal Zheng Zhou Univ. (Nat. Sci. Ed.), vol. 42, no. 6, pp. 66-71, 2010.

[13] T. Lu, H. Wang, and H. L. Yao, “K-nearest neighbor Chinese text categorization algorithm based on center documents,” Computer Engineering and Applications, vol. 47, no. 2, pp. 127-132, 2011.

[14] D. M. Deng, J. Z. Long, and X. Z. Yin, “A clustering algorithm based on structured Web document,” Journal of Central South University (Science and Technology), vol. 41, no. 10, pp. 1871-1875, 2010.

[15] Y. Z. Qu, W. Hu, and G. Cheng, “Constructing Virtual Document for Ontology Matching;” Proc. of 15th International Conference on World Wide Web (WWW 06), pp. 23-31, 2006.

 ISBN: 978-1-61208-419-0