http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet

Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (6): 1178-1185, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.1178.1185

Fuzzy Document Clustering using Weighted Conceptual Model

Shengli Song, Zengxin Guo and Ping Chen
Software Engineering Institute, Xidian University, No. 2 South Taibai Road, Xi'an, China

Abstract: Document clustering techniques mostly rely on single term analysis which can not reveal the potential semantic relationship between terms. To better capture the semantic subject of documents, this study proposes weighted conceptual model for document presentation. The new model divides the document concepts into centroid concepts and peripheral concepts due to their semantic relations to subject. The semantic similarity between two documents is calculated by centroid concepts and peripheral concepts respectively. A fuzzy semantic clustering method is put forward bases on the new semantic model. Experimental results show that the method enhances semi-structured document clustering quality significantly and outperforms K-Means and Fuzzy C-Means.

Key words: Text mining, document representation, conceptual graph, centroid concepts, peripheral concepts, semantic similarity

INTRODUCTION

Communication and knowledge sharing became very fast leads to vast amount of information spreading with a great speed. It is an important research topic to store, process and also obtain beneficial information from these data. If the data to be processed is either with no structure or semi-structured textual data then this process is called text mining (Weng and Lin, 2003). Text clustering is one of the important parts of the text mining system (Fu *et al.*, 2007). The goal for this kind of clustering is to divide a set of text documents into homogeneous groups and precisely distinguish truly relevant information.

Text clustering could be considered as the unsupervised learning process of organization textual documents into groups whose members are similar in some way, such as unobvious relations and structures can be revealed. Recently, text clustering has been receiving more and more attentions as an important method for textual information processing. It has been widely applied in document organization, automatic summarization, topic extraction and information filtering and retrieval (Wang *et al.*, 2007).

There are many well-known methods proposed for automatic text clustering. Generally, they can be classified into hierarchical methods and partitioning method. Hierarchical clustering methods proceed successively using a tree-like representation with different levels of granularity. These methods are good for visualizing and browsing through document collections. The main shortage is that the high complexity will increase according to the size of textual collection. On the other hand, partitioning methods compose the documents into several disjoint classes such that the documents in a class are more closely to one another than the documents in other classes (Li *et al.*, 2004). However, these methods have the shortcoming: clustering results are heavily dependent on the initial centroid seeds. Another limitation comes from the necessity of specifying a priori desired number of clusters to generate. Unfortunately, the optimum number is unknown a priori.

Recent studies for document clustering method use machine learning techniques, graphs-based method and matrix factorization based method. Machine learning based methods use semi-supervised clustering model with respect to prior knowledge and document's membership (Liu *et al.*, 2002; Ji and Xu, 2006; Zhang *et al.*, 2008). Graphs based methods model the given document set using an undirected graph in which each node represents a document (Hu *et al.*, 2008). Matrix factorization based method use semantic features of documents set for document clustering (Xu *et al.*, 2003; Xu and Gong, 2004; Park *et al.*, 2009).

Clustering by text concepts is useful to reduce the conceptual space for linking a query to relevant information. Text documents with similar concepts are grouped into the same cluster and clusters with similar concepts are nearby in the semantic space. This is the main difference between neural clustering and traditional statistical cluster analysis which only assigns documents to clusters by more items they shared but ignores the semantic relationship between clusters. The standard methods for text representation are mainly based on Vector Space Model (VSM) (Salton *et al.*, 1975). In the

Corresponding Author: Shengli Song, Software Engineering Institute, Xidian University, No. 2 South Taibai Road, Xi'an, China

model, each document is considered as a vector in the term space. So the similarity between two documents can be measured with vector distance (e.g., Euclidean distance, Manhattan distance). However, the clustering methods in existence mainly focus on similarity computation algorithm and text representation model. They rarely consider the document structure and the semantic hierarchy of concepts in document. And also, on vector representations of documents based on the bag-of-words model, text clustering methods tend to use all the words/terms in the documents after removing the stop-words. This leads to thousands of dimensions in the vector representation of document. However, only a very small number of words/terms in documents have distinguishable power on clustering documents. Hence, our aim was to consider different views onto the document. First, we use part-of-speech and WordNet semantic lexicon to decrease the number of words/terms. All the concepts extracted would be meaningful to the document subject and useful for clustering. Second, different semantic level at which text documents may be represented and from which clustering results are eventually derived.

In this study, we propose a Weighted Conceptual Model (WCM) to represent a document. WCM depicts the relationship among subject, centroid concepts and peripheral concepts of a text document. The similarity measure between documents is defined based on the semantic information contained in WCM. Based on the representation model, we put forward a Fuzzy Semantic Clustering method for document clustering.

WEIGHTED CONCEPTUAL MODEL

A wide range of different feature descriptions often represents documents. The most straightforward description of documents relies on term vectors. So document preprocessing has to be done, such as lexical analysis uses the word lexicon, elimination of stop-words, stemming and concepts selection based on the part of speech analysis of terms. Based on the concepts extracted, we can construct our WCM to represent the semantics of the document.

The WCM is defined as a directed acyclic graph which can be formal described as five tuples: G = (Subject, C, P, R, Φ). Where (1) Subject is a sentence describe the main idea of the documents; (2) $C = (c_1, c_2, ..., c_k)$ represents the centroid concepts of a document; (3) $p = (p_1, p_2, ..., p_m)$ represents the peripheral concepts of a document; (4) $R = \{(p_1, c_j) | \phi\ (p_1, c_j) \geq \beta\} \subseteq P \times C$ describes the semantic relationship from peripheral concept p_1 to c_j centroid concept and the semantic similarity more than threshold β; (5) $\Phi = \{\phi\ (p_i, c_j)\}$ is a weight value measure the semantic relevance from p_i to c_j. $\phi\ (p_i, c_j)$ is a function to compute the semantic similarity by using well-known semantic lexicon WordNet.

$$\begin{aligned}
&G' = (\text{subject}, C', P', R', \Phi') \\
&\text{where, } C' = \{c_1, c_2, c_k\}, \\
&\qquad P' = \{p_1, p_2, p_3, p_m\}, \\
&\qquad R' = \{(p_2, c_1), (p_m, c_1), (p_1, c_2), (p_3, c_k)\}, \\
&\qquad \Phi' = \{\phi\ (p_2, c_1), \phi(p_m, c_1), \phi\ (p_1, c_2), \phi\ (p_3, c_k)\}
\end{aligned}$$

Fig. 1: Weighted conceptual model definition of a sample document

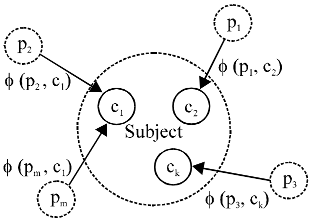

Fig. 2: Graphical representation of weighted conceptual model

```
Algorithm 1: Wieghted conceptual model representation construction algorithm
 Input: Insert (element): insert eliment into set.
         Posfs (term): Part of speech of term.
       $SIM_{WN}$ (term 1, term 2): WordNet similarity between term 1 and 2.
          Term-set, thershold, title
 Out put: G = (Subject, centroid-set, peripheral-set, relation, $\Phi$)
 Initial: Centroid-set = $\phi$, peripheral-set = $\phi$, similarity = 0, subject = title)
  for (term: Term-set)
  If (pofs(term) $\in$ {noun,verb})and (term ) $\notin$ candidate-set)
          If term $\in$ {Title, abstract, subtitles, key words}
              Insert (term,centroid-set)
           Else
              Insert (term, peripheral-set)
            end-if
      end-if
  end-for
 for (cterm: centroid-set)
      for (pterm: peripheral-set)
   Similarity = $SIM_{WN}$ (cterm, pterm)
           If similarity$\geq$thershold  {
           Insert ((cterm, pterm), relation)
              Insert (similarity, $\Phi$)
            }
           end-if
      end-for
 end-for
```

Fig. 3: WCM representation construction process

For a document with semi-structure which means the document has some structure information such as title, abstract or keywords, etc. which can be defined as Fig. 1. Figure 2 is the graphical representation for the document.

Figure 3 shows the algorithm of how to construct our WCM to represent a semi-structure document. Unlike traditional text representation models, WCM divides the concepts to two levels based on their importance to the subject. It also uses the semantic relevance to indicate the relationship the concepts from different levels.

SEMANTIC SIMILARITY

Semantic similarity is a generic issue in the areas of application areas of Artificial Intelligence and Natural Language Processing. Similarity between two words is often represented by similarity between the concepts related with the words. A number of semantic methods have been developed in literatures. Various similarity methods have proven to be useful in some specific applications. In general, the semantic similarity measures can be categorized into two groups: thesaurus-based methods and corpus-based methods. A detailed review on word similarity can be found from the research study by Li *et al.* (2003) and Rodriguez and Egenhofer (2003). Assuming a lexical taxonomy is constructed in a tree like hierarchy with a node for a concept, it has proven that the minimum number of edges connecting concepts c1 and c2 is a metric for measuring the conceptual distance of c1 and c2. The edge counting is useful for specific application with highly constrained taxonomies. However, lexical taxonomy may have irregular densities due to its broad coverage domain. Such a problem of no uniformity can be corrected by the utilization of depth in the hierarchy where word is found.

The basic idea of corpus-based methods is to define the similarity between two concepts as the maximum of the information context of the concept that subsumes them in the taxonomic hierarchy. The information content of a concept depends on the probability of encountering an instance of the concept in a corpus. However, for difficulties in acquiring appropriate corpus and the high computational complexity, its application has some limitations.

WordNet is an online semantic dictionary developed at Princeton by a group led by Miller (1995). It tries to make the semantic relations between word senses more explicit and easier to use. WordNet has 144,684 words and 109,377 synonym sets, named synsets. The synset reflects a concept in which all words have similar meaning. This design for concepts in WordNet is very similar to the concept organization in human natural language.

Similarity of two words is measured by the length of the shortest path between them in the hierarchical tree. In the case that a word is polysemous, multiple paths may exist between the two words. Yang and Powers (2005) proposed YP based on the assumptions that every single path in the hierarchical structures of WordNet 1) is identical and 2) represents the shortest distance between any two connected words. The similarity between two words w_i and w_j can be represented by the equation Eq. 1

- Words similarity:

$$\mathrm{Sim}_{WN}(w_i, w_j) = \begin{Bmatrix} \alpha_t \prod_{i=1}^{l-1} \beta_{t_i}; & 1 < \gamma \\ 0, & 1 > \gamma \end{Bmatrix} \qquad (1)$$

where, $0 \le (w_i,\ w_j) \le 1$; d is the depth of longest common subsequence; t represents the type of path which connects them; α_t represents their path type factor; β_t represents their path distance factor and γ represents an arbitrary threshold on the distance introduced for efficiency, representing human cognitive limitations. The values of α_t, β_t and γ have already been empirically tuned as 0.9, 0.85 and 12, respectively.

For the purposes of information retrieval or text mining, it is important to be able to measure the similarity of two texts represented with conceptual graphs. Poole and Campbell (1995) defined three kinds of similarity, i.e., surface, structure and thematic similarity. Surface or structure similarity is the similarity based on the matching of particular concepts or relations, while thematic similarity depends on the presence of particular patterns of concepts and relations. We will focus on sematic similarity in this study.

Since WCM consists of concepts and relations, we will define the similarity between WCMs based on the similarity between concepts of same level and the similarity between relations. Based on semantic similarity in WordNet, we can define the similarity measurement method of WCMs. Commonly, the centroid concepts of a WCM g can be used to express the subject which is the main idea of the corresponding document to g. We propose the WCM's similarity measure method as follow. Equation 2 defines the similarity between centroid concepts from two documents. Equation 3 defines the similarity between each simantic relationships from two documents. WCM similarity definition shows as Eq. 5.

- Centroid concepts similarity:

$$\mathrm{Sim}_{CC}(C_a, C_b) = \frac{1}{|C_a|} \sum_{i=1}^{|C_a|} \mathop{\mathrm{Max}}_{j=1}^{|C_b|} (\mathrm{Sim}_{WN}(c_i, c_j)) \qquad (2)$$

where, C_a and C_b are centroid concepts of two WCMs and suppose $|C_a| \le |C_b|$. For each centroid concept, $c_i \epsilon C_a$ get words similarity to each centroid concept $c_i \epsilon C_b$ and record the maximum value. The average score of is c_i used to indicate the value of centroid concepts similarity.

- Concept arcs similarity:

$$\mathrm{Sim}_{CA}(R_a,R_b)=\frac{1}{|R_a|}\sum_{i=1}^{|R_a|}\sum_{j=1}^{|R_b|}\mathrm{Equ}\,(r_i,r_j) \tag{3}$$

$$\mathrm{Equ}\,(r_i,r_j)=\left\{\begin{matrix}\phi_i, & r_i=r_j\\ 0, & r_i\neq r_j\end{matrix}\right\} \tag{4}$$

where, R_a and R_b are relationships of two WCMs and suppose $|C_a|\leq|C_b|$. For each relationship $r_i \in R_a$, if there is a relationship $r_j \in R_b$ equal to r_i (the two relationships have the same peripheral concept and centroid concept, respectively) we get the relevance from function Equ (r_i, r_j). The average score of r_i is used to indicate the value of concept arcs similarity.

- WCM semantic similarity:

$$\mathrm{Sim}_{WCM}(G_a,G_b)=\delta\cdot\mathrm{Sim}\,(C_a,C_b)+(1-\delta)\cdot\mathrm{Sim}\,(R_a,R_b) \tag{5}$$

The value of importance factor was δ set to 0.7 in present experiment to get better performance.

FUZZY DOCUMENT CLUSTERING

In text mining, clustering methods can be distinguished by regarding how they assign text document to clusters, i.e., what type of partitions they form. In classical cluster analysis each document must be assigned to exactly one cluster. These classical methods yield exhaustive partitions of the example set into non-empty and pair wise disjoint subsets. Such hard assignment of document to cluster can be inadequate in presence of documents that are equally distant to two or more clusters. Such special documents can represent hybrid-type or mixture objects which are equally similar to two or more clusters.

WordNet and ontology can be used to calculate the similarity of two words and the background knowledge can be used to integrate into the process of clustering text documents (Hotho *et al*., 2003, 2006). We use WordNet in our clustering process and choose the fuzzy clustering method similar to Zeng *et al.* (2004) and Doring *et al.* (2006).

Fuzzy clustering is an objective function based method to divide a document set into a set of groups or clusters. In contrast to hard assignment of document to clusters, fuzzy clustering offers the possibility to assign a document to more than one cluster, so that overlapping clusters can be handled conveniently and to assign it with degrees of membership. Each cluster is represented by a prototype which consists of a cluster center and maybe some additional information about the size and the shape of the cluster. The degree of membership, to which a document belongs to a cluster, is computed from relevance of the document to the cluster centers.

The relevance from document to cluster can be measured by semantic similarity based on ontology. By using the document's WCM representation, we can express the semantic similarity of two documents by their WCM similarity.

Therefore, similarity between documents and document to cluster can be calculated as Eq. 6 and 7.

- Documents similarity:

$$\mathrm{Sim}_{DD}(d_i,d_j)=\mathrm{Sim}_{WCM}(G_i,G_j) \tag{6}$$

- Document-cluster similarity:

$$\mathrm{Sim}_{DC}(d_i,\mathrm{Cluster})=\frac{1}{|\mathrm{Cluster}|}\sum_{d_j}\mathrm{Sim}\,(d_i,d_j) \tag{7}$$

where, G_i, G_2 is respectively the WCM representation of document d_i and d_j .

Formally, in fuzzy clustering we are given data set including N docuemts which is to be divided into K clusters where, K has to be specified by a user. The fuzzy clustering result is obtained by minimizing the objective function as Eq. 8.

$$J_2(U,V)=\sum_{i=1}^{N}\sum_{j=1}^{K}(u_{ij})^2\cdot\mathrm{Sim}_{DD}^{-2}(d_i,v_j) \tag{8}$$

Subject to$\forall_i$, $\forall_j$, $u_{ij}\geq 0$, $\forall i, \sum_{j=1}^{K}u_{ij}=1$.

where, $v_j \in \mathrm{Cluster}_j$ is the centroid document of cluster i. v_j can get from the equation Eq. 9.

$$v_j=d_j \tag{9}$$

$$\mathrm{St.min}\left\{\sum_{d_t\in\mathrm{Cluster}_i;d_j\neq d_t}\begin{bmatrix}\mathrm{Sim}_{DD}(d_j,d_t)-\\ \mathrm{Sim}_{DC}(d_j,\mathrm{Cluster}_i)\end{bmatrix}^2\right\}$$

Unfortunately, the objective function cannot be minimized directly. It is restricted by both the cluster prototypes and the membership degrees. Therefore, an iterative algorithm is used to optimize the cluster prototypes and the membership degrees, each time keeping the other set of parameters fixed at their current values. Differentiating the objective function (extended by Lagrange multipliers to incorporate the constraints) derives the update formulae with restrict to the parameters to optimize equal to zero. For the membership degrees one thus obtains the update equation.

$$u_{ij} = 1 \Bigg/ \sum_{k=1}^{K} \left[\frac{Sim_{DD}(d_i, v_k)}{Sim_{DD}(d_i, v_j)} \right]^2 \tag{10}$$

Subject to ∀i, ∀j, $Sim(d_i, v_j)>0$.

That is, the membership degrees basically reflect the relative squared semantic similarity of the document to the different cluster centers.

EXPERIMENTAL RESULTS

The 20-newsgroup data sets, collected by Lang, contain about 20,000 articles. Each newsgroup represents one class in the hierarchy structure. Each article is designated to one or more semantic categories and the total number of categories is 20. We choose 6 categories including 250 documents for the first experiment and 1000 documents for the second. All the documents selected randomly from the 20-newsgroup data sets and each document is about 5~6K.

We evaluate present approach by the recall R, precision P and F-measure. F combines R and P, in a single measurement as follows: $F = (\beta^2 + 1)\ PR/(\beta^2\ P + R)$. The parameter β influences how much to favour recall over precision. Researchers frequently report the F1 score of the system where, β = 1 weighting precision and recall equally. The article is correct clustered, if one of the original labels assigned by data set matches the cluster label. All the documents are respectively clustered by K-Means (KM), Fuzzy C-Means (FCM) and our Fuzzy Semantic Clustering (FSC) using WCM in each experiment.

The experiments revealed some interesting trends in terms of the clustering qualities of both datasets. Clearly, the results show the effectiveness of the WCM based approach for clustering. From the result data, we know that (1) the performance of FCM and FSC is better than KM clustering and (2) FSC is more stable than KM or FCM. Both of 48.45% of FCM and 49.2% of FSC are better than 33.43% of KM in Table 1. The same result is shown in Table 2 that 44.74% of FCM and 51.56% of FSC are better than 33.61% of KM. KM is not a stable clustering mothod. Because KM algorithm randomly selects document as initial cluster centers which may result in missing some clusters. There is no document grouped to the cluster of talk.politics.misc in Table 1 and alt.atheism in Table 2 neither. The differences between the highest and lowest value of FCM are about 0.6302 (0.6596-0.0294) in Table 1 and 0.4431 (0.6231-0.18) in Table 2. For the clustering result of FSC, the value are 0.1403 (0.4162-0.2759) and 0.4107 (0.5876-0.169). In addition, the experiments illustrate that 3 the performance of FSC is promoted when more documents clustered. Because more semantic clues are included in the process of measuring similarity between WCMs of documents. This affirms the assertion that the performance of the clustering algorithms is to proportionally improve as more semantic understanding of text content is considered. The results listed in Table 1 and 2 show the improvement in the clustering quality when more documents are included in the semantic-based similarity measure. Since we increase the size of data sets

Table 1: Clustering result using three methods (n = 250, K = 6)

Experiment 1	KM			FCM			FSC		
(Category)	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
Alt. atheism	0.14	1	0.2456	0.62	0.7045	0.6596	0.26	1	0.4127
Comp. graphics	0.04	1	0.0769	0.60	0.4839	0.5357	0.18	1	0.3051
Misc. forsale	1.00	0.1773	0.3012	0.02	0.0556	0.0294	0.74	0.3136	0.4408
Rec. motorcycles	0.02	1	0.0392	0.50	0.3247	0.3937	0.16	1	0.2759
Sci. space	0.12	1	0.2143	0.18	0.90	0.30	0.26	0.4483	0.3291
Talk. politics.misc	0.00	0	0	0.72	0.4045	0.5180	0.72	0.2927	0.4162
(SUM)	0.22	0.6962	0.3343	0.4788	0.4903	0.4845	0.3867	0.6758	0.4920

Table 2: Clustering result using three methods (n = 1000, K = 6)

Experiment 2	KM			FCM			FSC		
(Category)	Recall	Precision	F_1	Recall	Precision	F_1	Recall	Precision	F_1
Alt. atheism	0	0	0	0.1104	0.4865	0.180	0.6380	0.5445	0.5876
Comp. graphics	0.0238	1	0.0465	0.3214	0.2673	0.2919	0.1131	1	0.2032
Misc. forsale	0.0076	0.1667	0.0148	0.5039	0.3552	0.4167	0.9535	0.2426	0.3868
Rec. motorcycles	0.8923	0.2511	0.3919	0.6615	0.5890	0.6231	0.0923	1	0.1690
Sci. space	0.8232	0.6287	0.7129	0.1878	0.8947	0.3105	0.2265	0.7069	0.3431
Talk. politics.misc	0.0732	0.2143	0.1091	0.6829	0.3489	0.4619	0.5244	0.4155	0.4636
(SUM)	0.3033	0.3768	0.3361	0.4113	0.4903	0.4474	0.4265	0.6516	0.5156

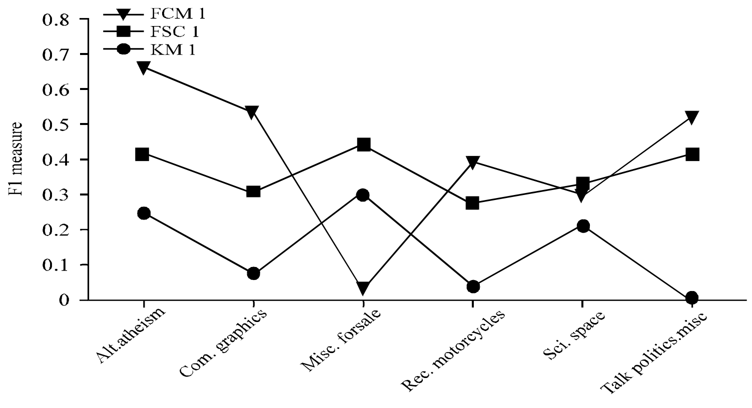

Fig. 4: The 250 documents clustering results from experiment 1

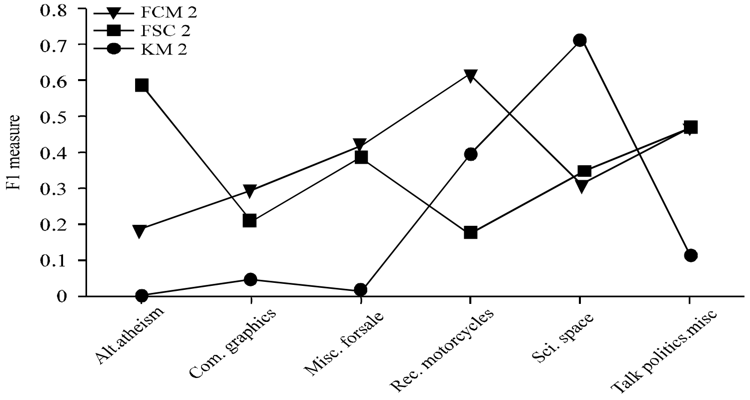

Fig. 5: The 1000 documents clustering results from experiment 2

from 250 to 1000, KM's performance is about 33% and has little change from the first experiment to the second. But FCM shows worse performance that is from 48.45 to 44.74%. The datasets size increment may augment the number of clusters which one documents may be grouped to. So the overall F-measure is decrease by about 4%. Contrary to the results of experiments by FCM, FSC shows better clustering performance from 49.2 to 51.56% according to the number of documents increase from 250 to 1000. We analyzed the reasons that more documents will introduce more semantic information. The information can raise the similarity of two document from one category in source. So, the larger size of the data sets, the semantic information will be more comprehensive and the better clustering performance will be shown. Other experiments proved that the more balanced number of documents in each cluster, the better clustering result will obtain.

To better understand the effect of the inclusion of semantics information when calculating similarity on clustering quality, we plot the clustering quality profile indices against the similarity options in Fig. 4 and 5. The plootted values are the F-measure of the different clustering algorithms. We can see that FSC outperforms FCM and HCM in both two experiments. It is the most stable method among them and the curve of it is smoother than others. The reason revealed is that the WCM is a stable method for document representation. It's easy to notice the enhancement of the FSC clustering as we use more documents.

By comparing the F-measure on all cases, FSC shows best performace and followed by FCM and KM sequentially. The semantic information used in FSC really play a role in the experiments.

DISCUSSION

Document representation is a fundamental issue for clustering and methods such as BOW, bags of phrases and n-grams have been widely investigated. Explicitly using external knowledge bases can assist generating concise representations of documents. The external knowledge includes WordNet, Domain Dictionary and Wikipedia, etc.

The Constructed Semantics Model (CSM) (Kwantes, 2005) is designed as a cognitive model to explain the emergence of semantics from experience. It operates by taking the term by document matrix (using barely log weighting) and multiplying it by its transpose. Consequently, terms do not have to appear together in order to be similar as is the case in the vector space model. The information of document we used in WCM is not the term sequences but the semantic relationship between concepts. The Conceptual Ontological Graph (COG) (Shehata *et al.*, 2007) analyzes terms on the sentence and document levels. The conceptual model can effectively discriminate between non-important terms with repect to sentence semantics and terms which hold the concepts that represent the sentence meaning. Present model is very like COG but different in concepts selection and document presentation. WCM is much simpler than COG model and presents the semantic relation of concepts by two different levels. The experimantal results shows the same effect of semantic information used in clustering. Statistical Topic Model (STM) (Chemudugunta *et al.*, 2008) provide a general data-driven framwork for automatic discovery of high-level knowledge from large collections of text documents. It can potentially discover a broad range of themes in a data set. But the model requires a hierarchy of human-defined semantic concepts to get better performance. Bag Of Concepts (BOC) (Huang *et al.*, 2009), a Wikipedia-based concept representation model is created by mapping the terms and phrases within documents to their corresponding articles in Wikipidia. Experimental results show that BOC model together with the semantically enriched document similarity measure outperform related approaches. We get the same conclusion through our experiments. A new Semantic Similairity Based Model (SSBM) proposed by Gad and Kamel (2009) utilizes WordNet to compute semantic similarity. It caputures the semantic similarity between documents that contain semantically similar terms but unnessarily syntactically identical. The model solves the ambiguity and synonymy problems and realizes the hidden similarities between documents due to the contribution of semantically similar terms as well as insensitivity to noisy terms. The semantic information in SSBM is also included in our WCM representation and we get better performance than SSBM.

CONCLUSION

In this study, we proposed Weight Conceptual Model for document representation. In this model, the concepts were extracted by the semantic similarity to the subject of document. The dimension of conceptual model is lower than vector space. The concepts used for describing the subject are divided into two different levels which are centroid concepts and peripheral concepts. The semantic relations between WCM are concerned. We also demonstrated the fuzzy semantic clustering approach based on WCM. Experiments on textual data sets show that fuzzy semantic clustering with WCM yields a better performance.

However, some defects, like the time performance of the semantic relevance measure with WordNet, are still in existence and need to do some future works.

ACKNOWLEDGMENTS

This study was supported by the Chinese Defense Advance Research Program of Science and Technology (513060601) and the Fundamental Research Funds for the Central Universities of China (K50510230002).

REFERENCES

Chemudugunta, C., P. Smyth and M. Steyves, 2008. Combining concept hierarchies and statistical topic models. Proceeding of the 17th ACM Conference on Information and Knowledge Mining, Oct. 26-30, Napa Valley, California, USA., pp: 1469-1470.

Doring, C., M.J. Lesot and R. Kruse, 2006. Data analysis with fuzzy clustering methods. Comput. Stat. Data Anal., 51: 192-214.

Fu, Y., D. Yang, T. Wang, A. Gao and S. Tang, 2007. A new text clustering method using hidden markov model. Nat. Language Processing Inform. Syst., 4592/2007: 73-83.

Gad, W.K. and M.S. Kamel, 2009. New semantic similarity based model for text clustering using extended gloss overlaps. Machine Learning Data Mining Pattern Recognition, 5632: 663-677.

Hotho, A., S. Staab and G. Stumme, 2003. Wordnet improves text document clustering. Proceedings of the SIGIR 2003 Semantic Web Workshop, (SWW'03), Toronto, Canada, pp: 541-544.

Hotho, A., S. Staab and G. Stumme, 2006. Ontologies improve text document clustering. Proceedings of the 3rd IEEE International Conference on Data Mining, Nov. 19-22, Melbourne, Florida, USA., IEEE Press, pp: 541-544.

Hu, T., H. Xiong, W. Zhou, S.Y. Sung and H. Luo, 2008. Hypergraph partitioning for document clustering: A unified clique perspective. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, Singapore, pp: 871-872.

Huang, A., D. Milne, E. Frank and I.H. Witten, 2009. Clustering documents using a wikipedia-based concept representation. Adv. Knowledge Discovery Data Mining, 5476: 628-636.

Ji, X. and W. Xu, 2006. Document clustering with prior knowledge. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 6-11, Seattle, Washington, USA., pp: 405-412.

Kwantes, P.J., 2005. Using context to build semantics. Psychonomic Bull. Rev., 12: 703-710.

Li, T., S. Ma and M. Ogihara, 2004. Document clustering via adaptive subspace iteration. Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, July, 25-29, South Yorkshire, UK., pp: 218-225.

Li, Y.H., Z. Bandar and D. McLean, 2003. An approach for measuring semantic similarity using multiple information sources. IEEE Trans. Knowledge Data Eng., 15: 871-882.

Liu, X., Y. Gong, W. Xu and S. Zhu, 2002. Document clustering with cluster refinement and model selection capabilities. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Aug. 11-15, Tampere, Finland, pp: 191-198.

Miller, G.A., 1995. WordNet: A lexical database for English. Commun. ACM, 83: 39-41.

Park, S., D.U. An, B.R. Cha and C.W. Kim, 2009. Document clustering with cluster refinement and non-negative matrix factorization. Proceedings of the 16th International Conference on Neural Information Processing, Dec. 1-5, Bangkok, Thailand, pp: 281-288.

Poole, J. and J.A. Campbell, 1995. A Novel algorithm for matching conceptual and related graphs. Conceptual Struct. Appl. Implementat. Theory, 954: 293-307.

Rodriguez, M.A. and M.J. Egenhofer, 2003. Determining semantic similarity among entity classes from different ontologies. IEEE Trans. Knowledge Data Eng., 15: 442-456.

Salton, G., A. Wong and C.S. Yang, 1975. A vector space model for automatic indexing. Commun. ACM, 18: 613-620.

Shehata, S., F. Karry and M. Kamel, 2007. A concept-based model for enhancing text categorization. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA., Aug. 12-15, pp: 629-637.

Wang, F., C. Zhang and T. Li, 2007. Regularized clustering for documents. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 23-27, Amsterdam, Netherlands, pp: 95-102.

Weng, S.S. and Y.J. Lin, 2003. A study on searching for similar documents based on multiple concepts and distribution of concepts. Expert Syst. Applications, 25: 355-368.

Xu, W. and Y. Gong, 2004. Document clustering by concept factorization. Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval, July 25-29, South Yorkshire, UK., pp: 202-209.

Xu, W., X. Liu and Y. Gong, 2003. Document clustering based on non-negative matrix factorization. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, July 28-Aug. 1, Toronto, Canada, pp: 267-273.

Yang, D.Q. and D.M.W. Powers, 2005. Measuring semantic similarity in the taxonomy of WordNet. Proceedings of the Australasian Conference on Computer Science, (ACSC'05), Australian Computer Society, Inc., pp: 315-322.

Zeng, H., Q. He, Z. Chen, W. Ma and J. Ma, 2004. Learning to cluster web search results. Proceeding of 27th Annual International ACM SIGIR Conference on Research and Development in Informing Retrieval, Jul. 25-29, Sheffield, South Yorkshire, UK., pp: 210-217.

Zhang, X., X. Hu and X. Zhou, 2008. A comparative evaluation of different link types on enhancing document clustering. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, Singapore, pp: 555-562.