I.J.Modern Education and Computer Science, **2009, 1, 35-44**
Published Online October 2009 in MECS (http://www.mecs-press.org/)

An Improved Text Clustering Method based on Hybrid Model

Jinzhu Hu
Department of Computer Science of HuaZhong Normal University, Wuhan, China
Email:xiongchunxiu@126.com

Chunxiu Xiong ,Jiangbo Shu,Xing Zhou,Jun Zhu
Department of Computer Science of HuaZhong Normal University, Wuhan, China
Email:xiongchunxiu@126.com

***Abstract*— According to the high-dimensional sparse features on the storage of textual document, and defects existing in the clustering methods or the hybrid methods which have already been studied by now and some other problems. So an improved text clustering method based on hybrid model, that is a text clustering approach (short for TGSOM-FS-FKM) based on tree-structured growing self-organizing maps (TGSOM) and Fuzzy K-Means (FKM) is proposed. The method has optimized the clustering result through three times of clustering. It firstly makes preprocess of texts, and filters the majority of noisy words by using an unsupervised feature selection method. Then it used TGSOM to execute the first clustering to get a rough classification of texts, and to get the initial clustering number and each text's category. And then introduced LSA theory to improve the precision of clustering and reduce the dimension of the feature vector. After that, it used TGSOM to execute the second clustering to get more precise clustering results, and used supervised feature selection method to select feature items. Finally, it used FKM to cluster the result set. In the experiment, it remained the same number of feature items and experimental results indicate that TGSOM-FS-FKM clustering excels to other clustering method such as DSOM-FS-FCM, and the precision is better than DSOM-FCM, DFKCN and FDMFC clustering.**

***Index Terms*—tree-structured growing self-organizing maps; Fuzzy K-Means; text clustering; text clustering flow model**

I. INTRODUCTION

With the prompt expansion of the WWW, as well as the rapid growth of a variety of textual documents at home and abroad, clustering textual documents is useful for the user's querying for the information of the textual documents, so the study of text clustering has paid attention to the related members of the public. So far, a variety of text clustering methods have been developed, but they have some defects in some aspects, therefore, improving the text clustering algorithms according to their shortcomings can achieve better clustering results.

Text clustering methods can be divided into division-based clustering methods, level-based clustering method, density-based clustering method, model-based clustering method and grid-based clustering method as well as other methods. But the most often used methods are K-Means、Single-Link、DBSCAN[7] and SOM method, and each has its own advantages and disadvantages. K-Means method is the most representative algorithm of the division-based clustering methods, the advantage of its process is very simple, and its complexity is relatively low, and its disadvantage is that it is seriously influenced by the initial cluster centers and easily affected by the outliers, and not easy to discover the concave shaped cluster, it also needs to specify the cluster numbers, at last the algorithm can only reach the local optimal. Therefore, the algorithm should first design a method which can determine the suitable initial cluster centers and cluster numbers, and also needs to execute a special method which can screen off the isolated points to reduce the impact of isolated points, finally, executing the clustering methods of different level to achieve the global optimum; Single-Link algorithm is derived from the agglomerate level-based clustering method, its advantage is that it can display the textual data at different levels in order to help people efficiently browse the textual data sets of large-scale, but also its shortcomings can not be ignored, for the time complexity is very high and may produce chained cluster, so to the algorithm should design a method which can reduce the time complexity step by step and gradually eliminate the possibility of chained cluster; DBSCAN algorithm is a density-based clustering algorithm, its advantage is it ables to handle clusters of arbitrary shape, the drawback is that it is too sensitive to two parameters ε and MinPts which need users to customize, therefore, to the algorithm, it should design a method which can easily determine the suitable parameters ε and MinPts; SOM algorithm is a model-based clustering method, the advantage is that it is an unsupervised clustering algorithm, and it is able to

Manuscript received January 4, 2009; revised June 5, 2009; accepted July 3, 2009.
Supported by the Priority Projects of the Key Research Base of Ministry of Education (Grant No.07JJD740063), the Fund of Hubei Physical Science (Grant No.2007AA101C49）

cluster input mode automatically, the shortcomings of the algorithm is that it is distorted when the dimension is reduced, and need to output great numbers of weight vector and a huge neuron number, if the amount of data is large, the study efficiency will be reduced. Therefore, to the algorithm, it should design a method which can prevent its distortion, reduce the scale of the output weight vector, as well as reduce the number of neurons in the premise of the efficiency of the algorithm does not affect.

The text clustering methods mentioned above are a hard category that is precise classification to text, in the process of clustering, it needs to determine the exact type of the text. However, not each of the textual documents belongs to only one category, but more in reality, it needs to analyze specifically to the specific textual documents in the process of the classification. According to the ambiguity of the text scope, it is better to use fuzzy clustering approach to complete the fuzzy text clustering. Because the fuzzy k-means can change the precise classification to fuzzy classification, and redefine the criterion function of the clustering through the introduction of the membership function, fuzzy k-means (FKM) can easily implement the fuzzy text clustering. But FKM algorithm is developed in the basis of the k-means method, so it needs to value the cluster number at first, which is the value of the parameter k, then it can combine with another clustering algorithm to achieve efficient text clustering. So far the main algorithms which can obtain the parameter k are dynamic self-organizing maps (DSOM) method, SOM method, Growing Self-Organizing Maps (GSOM) method, Tree-structured Growing Self-Organizing Maps (TGSOM)[5] method and so on. Although the SOM-based text processing algorithm is suitable for large-scale textual document clustering, it can't achieve extracting cluster result automatically, so it has some limitation. Although the DSOM algorithm can achieve extracting cluster result automatically, but it obtains the clustering result only through once clustering, therefore, the process of noisy words, although the feature selection methods can remove most of the words, there is a sharp decline in the text clustering performance. Consequently, it is better to search for another processing method to strengthen the text clustering performance. Using Latent semantic analysis[4] (LSA) theory to strengthen the relationship between semantics, and greatly reduce the vector space so as to achieve a higher rate of text clustering and reduce the noise factors, after that, it can take the second cluster to improve the clustering accuracy.

Obviously, DSOM algorithm can not meet the requirements of the second cluster, but the TGSOM can cluster more. The advantage of the Dynamic Growing Self-Organizing Maps (GSOM) is that the visualization effect of the clustering results is very good, but the disadvantage of which is that, it can't generate new nodes at the right time and in the right position which can cause low efficiency to the implementation of the method. Because TGSOM inherited the basic idea of Kohonen SOFM, and used a flexible tree structure, it can generate new modes at the right time and in the right position, so its network structure was dynamically generated, and overcame the limitation of the immobilized of the SOFM network structure, at the same time the efficiency of the implementation of TGSOM method is significantly higher than GSOM. The tree structure used by TGSOM is similar to the Self-creation and self-Organizing Neural Networks model (SCONN), but TGSOM algorithm has used the spread factor (SF) to control the growth of the network, and realized the hierarchical clustering, so the visibility of the training results is obviously better than SCONN.

To sum up, TGSOM is more suitable for large-scale text clustering, TGSOM combines with the FKM to the text clustering is a highly efficient algorithms.

In text clustering, the most commonly used description method to the text is Vector Space Model (VSM), which regards each word as a dimension of the feature space coordination, and each textual document as a vector of the feature space. Although the description is simple and direct, for the words of the each textual document is different and the count of the words have great difference, it makes the text vector space high-dimensional and sparse, text clustering of a significant decline in performance, and high complexity. These clustering results should be improved so as to satisfy the user. So the main process is to reduce the dimension of the text vector space, the improve way is to adopt feature selection to reduce dimension. So far, it has two feature selection, that is supervised feature selection and unsupervised feature selection. Supervised feature selection needs to calculate the relationship between class and feature to select a subset, however, the beginning described text vector space does not have the class information, so the feature selection can not be used in text clustering directly, but it can select the subset which has the most distinguish ability, therefore, it can use the characteristic of the supervised feature selection after the text vector space has the class information. But the unsupervised feature selection is just contrary to the supervised feature selection, it is difficult to select the subset which has the most distinguish ability, but it can be used in case in which does not have class information. Therefore, we can adopt unsupervised feature selection before the first cluster, and then adopt the supervised feature selection after producing type information will make text clustering produce a better result.

After the above analysis, we put forward a novel text clustering flow model (TCFM), which firstly executes unsupervised feature selection to the original data, and delete the noise words. Secondly, executes the first cluster to determine the cluster number and the documents type information. Thirdly, executes LSA to clear more noise words, strengthen semantics relationship, reduce the vector space, which improves efficiency and gets more accurate clustering results for the second cluster. Fourthly, it executes supervised feature selection to further reduce text space dimension

and select the subset which has the most distinguish ability. At last, through the third cluster of FKM to improve the cluster result. A novel text clustering approach (short for TGSOM-FS-FKM) based on TGSOM and FKM is proposed.

The organizational structure of the article as follows: The second part introduces the text pre-processing methods; the third part describes the TCFM; and the fourth part introduces the TGSOM-FS-FKM clustering method in detail; at last, summarizing the full text.

II. TEXT PREPROCESS

For Chinese text, the text pre-processing is mainly about the segmentation of the textual documents, removing stop words, counting the word frequency, calculating the weight and generating VSM, it can also be improved according to the need.

A. The Description of Text Clustering

Suppose that text-based data set $DS = \bigcup d_i$ (where i = 1 ... N, N is the number of the textual document in data set DS), d_i is the textual document in data set DS, the task of text clustering is to separate data set DS into the set of cluster C_j that is $DS = \bigcup C_j$ (where j = 1 ... M, M is the number of the cluster), the requirements as the formula (1) and (2) follows:

$$\max(sim(d_i, C_j)) \quad , \quad if \quad d_i \in C_j \tag{1}$$

$$\min(sim(d_i, C_j)) \quad , \quad if \quad d_i \notin C_j \tag{2}$$

The purpose of text clustering is to make the similarity of the relevant documents bigger and that of the unrelated documents smaller as far as possible, and according to the document's similarity, divide a large text set into many subsets called clusters. The similarity among the documents is large in a cluster and small in different clusters.

B. Text Representation

The description which widely used of the text data is the VSM proposed by the Professor Salton. The model for that text expressed as $C_i = (c_{i1}, w_{i1}; c_{i2}, w_{i2}; ...; c_{ij}, w_{ij}; ...c_{in}, w_{in})$, where c_{ij} is feature word, w_{ij} is a weight of c_{ij} in document C_i. The similarity between two documents is measured by the traditional Cosine distance, as following formula (3):

$$sim(C_i, C_j) = \frac{\sum_{l=1}^{n} w_{il} \cdot w_{jl}}{\sqrt{\sum_{l=1}^{n} w_{il}^2} \cdot \sqrt{\sum_{l=1}^{n} w_{jl}^2}} \tag{3}$$

Where the weight w_{ij} is calculated by the TF-DF formula, as follows formula (4):

$$w_{ij} = \frac{tf(c_{ij}, C_i) \times \log(n/n_t + 0.01)}{\sqrt{\sum [tf(c_{ij}, C_i) \times \log(n/n_t + 0.01)]^2}} \tag{4}$$

Where $tf(c_{ij}, C_i)$ is the frequency of c_{ij} in C_i, n is the total of the text, n_t is text number of c_{ij} appears in the practice text.

C. Feature Selection

Due to the particularly high dimension of the text vector, the complexity of the corresponding clustering method is also very high, at the same time it will lead into the interference, so it needs to filter the feature words, delete stop words, and remove noise words. Now we'll introduce two kinds of supervised learning feature selection methods and three kinds of unsupervised learning feature selection methods respectively.

supervised learning feature selection methods

Information Gain(IG)

Information Gain(IG) expresses the measurement of the amount of the information which contained in the document. The improvement of the IG of the text which will be clustered as the following formula(5):

$$IG^*(c) = \log_2(n_c + 0.01) \times IG(c) \tag{5}$$

Where n_c is the text numbers in which c appeared in the training text set, IG(c) is the information gain of the text which will be clustered as the following formula(6):

$$IG(c) = H(C) - H(C/c)$$

$$= -\sum_{C_i \in C} \frac{c_n(C_i)}{\sum_{C_i \in C} c_n(C_i)} \log_2 \frac{c_n(C_i)}{\sum_{C_i \in C} c_n(C_i)} + \sum_{C_i \in C} \frac{c_n(C_i/c)}{\sum_{C_i \in C} c_n(C_i/c)} \log_2 \frac{c_n(C_i/c)}{\sum_{C_i \in C} c_n(C_i/c)} \tag{6}$$

Where c is the word, C is the textual document set, c_n is the count of the different words in the document C_i. The improved $IG^*(c)$ has reflected the category capacity of c to the aggregation of the text which will be clustered, that is to say,

The greater value of $IG^*(c)$, the more information of c contains, the more information which can be provided to the cluster, so taking c as the feature word, the clustering will be more determined.

χ^2 statistics (CHI)

According to the literature [3], the inadequacies of χ^2 statistics (CHI) is that, the feature word contribute little to the classification when the frequency of the feature word appears higher in other class but lower in the designated class, so it should be deleted, but the calculated value of χ^2 statistics is very high, which can not meet demand. Therefore, in this paper, supervised

feature selection method applies χ^2 statistics which has been improved. Supposed that practice cluster set $D_i = (d_{i1},...,d_{ij},...,d_{in})$, and the improved formula as the following formula (7):

$$\chi^2(c,d) = \frac{(f_{cd} \cdot f - f_c \cdot f_d)^2}{(f_{cd} + f_c)(f + f_d)} \times \frac{\alpha + \beta + \gamma}{3} \quad (7)$$

Where $\alpha = \sqrt{\sum_{j=1}^{n} f_{ij}^2}$, $\beta = \frac{(n_i - t_i)^2}{t_i}$, $t_i = \frac{\sum_{i=1}^{n} n_i}{n}$, $\gamma = n_i$, f_{cd} is the frequency of the document which contains c and belongs to cluster d, f_c: contains c but not belongs to d, f_d : don't contain c but belongs to d, f : don't contain c as well not belong to d. α is the frequency of c in d, β is the concentration of c in d, γ is dispersity. f_{ij} is the frequency of c in text c_{ij} ($1 \le j \le n$), n_i is the number of document which contains c in D_i, n is text cluster's total number.

ReliefF Algorithm

ReliefF algorithm has a good performance, and it is the most representative algorithm of the supervised feature selection. The basic idea of the ReliefF algorithm is that by selecting a sample from the sample set randomly, and then learning some neighbors of the sample, at the same time calculating the weight of each feature word, repeating the above operations and updating the weight of each feature word, after several circulation, selecting several feature words which has the larger weights.

The steps of the ReliefF algorithm as follows:

Supposed that the sample set $S = \{s_1, s_2, \cdots, s_i, \cdots, s_n\}$, where the sample $s_i = [s_{i1}, s_{i2}, \cdots, s_{ij}, \cdots, s_{im}]$, s_{ij} is the jth feature value in the ith sample (where i=1, …,n; j=1, …,m), $w_i = [w_{i1}, w_{i2}, \cdots, w_{ij}, \cdots, w_{im}]$ represents the weight vector of the mth feature word in the ith sample, that is to say, w_{ij} is the weight vector of the jth feature word in the ith sample.

S1: First of all, w_{ij} =0（1≤i≤n,1≤j≤m）, that is the initialized weight of each feature word is the same, that is to say, at the very beginning the importance of all of the feature words is equal;

S2：Selecting a sample s_i randomly;

S3：Finding out k samples s_l' (l =1, …,k) which have the same class and the nearest distance with the sample s_i;

S4：Finding out k samples s_{cl}'' (l =1, …,k;c≠ class(s_i)，class(s_i) expresses the class of the sample s_i) which have the nearest distance but different class with the sample s_i in the sample set ;

S5：Updating the weight vector, its calculation method as the following formula (8):

$$w_{ij} = w_{ij} - \frac{\sum_{l=1}^{k} dif(f_{ij}, s_i, s_l')}{(r \times k)} + \frac{\sum_{c \ne class(s_i)} [\frac{p(c)}{1 - p(class(s_i))} \sum_{l=1}^{k} dif(f_{ij}, s_i, s_{cl}'')]}{(r \times k)} \quad (8)$$

Where $dif(f_{ij}, s_m, s_n') = \frac{|v(f_{ij}, s_m) - v(f_{ij}, s_n')|}{\max(f_j) - \min(f_j)}$, that is, the function $dif(\cdot)$ expresses the differences of two samples about a feature word, f_{ij} is the jth feature word in the ith sample s_i, $v(f_{ij}, s_i)$ is the value of the jth feature word in the sample s_i, $\max(f_j)$ is the largest value of the jth feature word of all of the samples in the sample S, and $\min(f_j)$ is the smallest value of the jth feature word of all of the samples in the sample set S, r is the iteration number of the algorithm, $p(c)$ is the probability which the c class appears, as the following formula (9):

$$p(c) = \frac{\sum_{i=1}^{M} x_i}{\sum_{j=1}^{N} y_j}, (x_i = \begin{Bmatrix} 1, s_i \in c \\ 0, s_i \notin c \end{Bmatrix}; y_j = \begin{Bmatrix} 1, s_j \in S \\ 0, s_j \notin S \end{Bmatrix}) \quad (9)$$

Where M is the total number of the c class's sample, N is the total number of the sample in data set S, that is $p(c)$ is the proportion of the total number of the c class's sample in the total number of the sample in data set S;

S6：Repeat the steps S2~S5, the iteration number is r.

Unsupervised learning feature selection method

Document Frequency(DF)

Document Frequency is the simplest feature selection method, which refers to the document numbers in the

whole of the data set which contains the feature word. The calculation method as the following formula (10):

$$DF = \frac{\sum_{i=1}^{N} x_i}{\sum_{j=1}^{N} y_j}, (x_i = \begin{Bmatrix} 1, word \in T_i \\ 0, word \notin T_i \end{Bmatrix}; y_j = \begin{Bmatrix} 1, T_j \in DS \\ 0, T_j \notin DS \end{Bmatrix}) \quad (10)$$

Where DS expresses the whole date set, $word$ is the feature word which needs to solve the Document Frequency now, T_i and T_j expresses the ith and the jth text in data set DS respectively.

The largest advantage of the Document Frequency is that the speed is very fast, for the relation between the complexity of which and the text numbers is linear relationship, classification results are very perfect especially when those rare words which are deleted are exactly the noise words. Consequently, the feature selection method is very suitable for the feature selection of the ultra-large-scale data sets. However, its drawback is that the perfect classification results has a prerequisite, that is, to a category, the word of which the frequency is too high or too low is meaningless, but in most cases, there are many key words which appear less in a document, if which are deleted as noise words, it would reduce the accuracy of the cluster.

Term Strength（TS）

The calculation of the Term Strength is probability of a word appears in one of the related documents when it appears in the other document. The calculation method as the following formula (11):

$$TS(word) = p(word \in s_i \,|\, word \in s_j), s_i, s_j \in DS, similarity(s_i, s_j) > \alpha \quad (11)$$

Where $word$ is a feature word, s_i, s_j expresses the related document, DS is the data set, $similarity(s_i, s_j)$ denotes the similarity between s_i and s_j, α is the similar threshold.

Term Strength thinks of a word is more important if it appears more in the related texts but less in irrelated texts.

Word contribution(WC)

Word contribution (WC) is the degree of a word contribute to the similarity of an entire textual data set. The calculation as the following formula (12):

$$WC(wd) = \sum_{i,j \cap i \neq j}^{N} w(wd, s_i) \times w(wd, s_j) \quad (12)$$

Where wd denotes a feature word, N is the total of the texts which contained in the textual data set DS, s_i, s_j denotes two texts, $w(wd, s)$ expresses the weight of the feature word wd in the text s.

III. TCFM

The characteristic of the textual data is high-dimensional sparse, so far, there is no method that can reduce the dimension very good at the same time maintaining the original function of the cluster or even improving it, but it can apply several suitable methods to delete noise words at the same time to improve the function of the cluster, so basing on the idea, the paper proposed a novel text clustering flow model(TCFM) through several times cluster and combining feature selections and LSA, the effectiveness of which has been approved by the final experiment. The TCFM first to preprocess of the textual data set, after that, filtering the majority of words which are useless or have little use by using unsupervised feature selection method. Then executing the first rough clustering by using a little spread factor, in order to delete the noise word further, applying the LSA theory to process the above texts, and then executing the second cluster to the texts by using a bigger spread factor, after having the clustering information, it can apply the supervised feature selection method to delete the feature words further which have little importance, at last, executing the fuzzy clustering to the texts to improve the accuracy of the clustering. The flow chart of TCFM is shown in Fig. 1.:

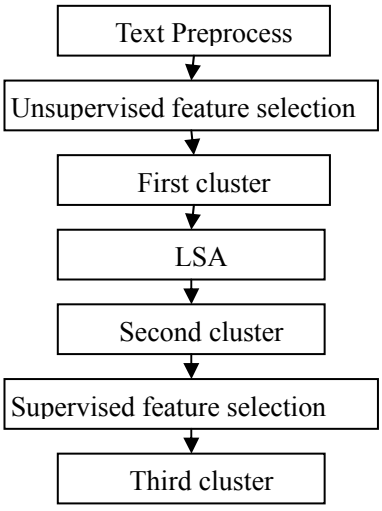

Figure 1. TCFM

We will introduce each process functions in detail, and explain its reasonable effectiveness.

A. Text Preprocess

Text preprocess includes segmenting words, deleting stop words, calculating term weight and generating VSM, etc.

So far, there are some Word Segmenter such as the good and fast word segmenter, Yard Chinese word segmenter, simple Chinese word segmenter (SCWS), Abot Chinese word segmenter(ACWPS), the third

generation of intelligent segmenter(3GWS), Chiru segmenter and ICTCLAS Chinese word segmenter.

The good and fast word segmenter adopted the technical line of character forms word and discriminate model based on classification, so it overcame the inherent defects of the traditional generated modeling theory, and has the perfect theoretical foundation, high segmentation precision, fast processing speed, and high learning efficiency, rapid deployment of the new application and new languages, and so on, it was mainly used in word processing, information retrieval, automatic abstracting, information extraction, monitoring public opinion, information filtering, competitive intelligence analysis, machine translation, speech processing and many other fields.

Yard Chinese word segmenter has only appears version 0.1, and the version has many places which need to be improved, for example, it can not segment the Chinese people names, place names, numbers and English, moreover, there are many problems in the division of ambiguous sentences.

The full name of SCWS is Simple Chinese Words Segmentation,

It is a mechanical Chinese word segmentation engine based on word frequency dictionary, and can segment a whole section of Chinese characters into words nearly correctly.

Abot Chinese word segmenter(ACWPS) is a natural language processing systems (NLP) which can separate a section of text into phrase by conventional automatically.

The 1st Generation Word Segmenter is a machinery rules method based on the linguistics knowledge, such as the greatest match and least segmentation methods and the error-driven mechanism; the 2nd Generation Word Segmenter is a machine learning method based on the large-scale corpus, such as the N-language model, the channel-noise model and the greatest expectations, etc. But the two systems neither have a relatively unified model framework which integrates the segmentation algorithm, segmentation disambiguation and unknown word recognition organically, nor have a unified evaluation system which can evaluate the segmentation results. However, the 3rd Generation Word Segmenter(3GWS) has integrated the latest research results in the statistical methods, semantic networks, model reasoning and language evolution and other areas to introduce an effective segmenter, of which mainly functions are the smart Chinese word segmentation, named entity and new word recognition, part of speech tagging and supporting the user-defined dictionary, but the disadvantage of the segmenter is that the accuracy of the part of speech tagging is low.

Chiru segmenter applied the structure of a Chinese Dictionary based on the dual-array Trie, and its segmentation algorithm is grid algorithm based on the local ambiguous word, while the shortest path search algorithm is based on the Fibonacci heap. Through these methods, it reduced the time complexity of the segmentation effectively, and it is mainly to solve the errors of the vocabulary segmentation in the special areas such as the agriculture area.

The main function of Institute of Computing Technology, Chinese Lexical Analysis System (ICTCLAS) contains Chinese segmentation, part of speech tagging, named entity recognition and new words recognition, at the same time supporting the user dictionaries, also it is the world's most popular open-source Chinese word segmenter, and has acquired the highest grade in the first comprehensive international segmentation contest, and in the national 987 evaluation. Consequently, the paper applied the ICTCLAS Chinese Segmenter.

B. Unsupervised Feature Selection

If unsupervised feature selection methods remove more words, it will reduce the text clustering performance. However, the biggest advantages of the document frequency (DF) is its high speed and low time complexity, suitable for ultra-large-scale data sets of feature selection, so make use of the advantages of DF and use the feature selection methods based on document frequency and feature similarity (DFFS)[2]. DFFS need three times feature filter, firstly, filter stop words to text data after automatic segmenting, at the same time, count word frequency and document frequency. Secondly, make use of DF to reduce most unrelated feature words. At last, use feature similarity to select feature subset and reduce some feature words, as shown in Fig. 2.

Figure 2. three times feature filter of DFFS

The steps of the DFFS method as follows:

S1：Segmenting the data sets automatically;

S2：Deleting the stop words according to the table of the stop words (the 1st feature filter), at the same time counting the word frequency and the document frequency;

S3：Deleting the words of which the values are too high or too low(the 2nd feature filter);

S4：Calculating the relevance of the feature words according to the method of the feature similarity, and selecting the feature subsets(the 3rd feature filter);

S5：Calculating the weights of each feature word.

C. First Cluster

Selecting the appropriate spread factor SF, which is very little to execute TGSOM algorithm, that is the 1st rough cluster, and obtaining the cluster number preliminary.

D. LSA

How to improve LSA according to literature [4]. Firstly, change text data expressed by VSM into a $m \times n$ matrix term A after the first cluster, n is the total number of different feature words, m is the dimension

of VSM, A can be expressed as the following formula (13):

$$A = [w_{ij}]_{m \times n} \quad (13)$$

Where w_{ij} is the weight expressed in formula (8). Secondly, execute singular value decomposition, A can be resolved as the following formula (14):

$$A = L \Lambda R^T \quad (14)$$

Where L is $m \times l$ matrix, R is $n \times l$ matrix, LSA technique just takes k singular values of matrix Λ and others set 0, where k is the cluster number after the first cluster.

E. Second Cluster

Execute TGSOM algorithm and use a higher SF value, it is not only fast but could also improve the cluster accuracy.

F. Supervised Feature Selection

The second cluster has provided cluster number for supervised feature selection, the feature selection can further reduce text words which can't distinguish text type and improve cluster accuracy.

G. Third cluster

The third cluster use fuzzy cluster. After the above steps, only a few text words which can distinguish text type well remained, and we got the cluster number, so it's better to use the effective method of FKM cluster.

IV. A NEW TEXT CLUSTERING METHOD BASED ON TGSOM AND FKM

The flow of method as mentioned above, we now introduce two clustering methods as follows:

A. TGSOM

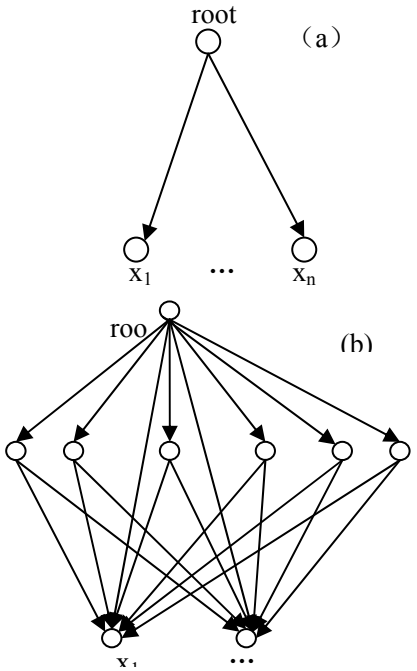

Figure 3. TGSOM structure of the model

The structure of TGSOM is composed of two parts, which are the input layer and the output layer. Fig.3 shows its graph. The initial state of web shows in Fig.3(a), and there is only a root node in the compete layer and n nodes $x_1 \sim x_n$ in input layer. Fig.3(b) shows the state of the web that growing to 6 nodes, and these 4 nodes composed a 2-D tree structure which make 'root' as its root node. The 6 nodes make complete interconnection with $x_1 \sim x_n$ that in input layer, and the weight of the node will be self-organized in the growing process.

Definition 1: A node is called the optimal match node if it is lies in compete layer and has the closest distance with the input vector v of the web, denoted as b. That is $\| v - w_b \| \leq \| v - w_n \|$, $\forall n \in N$, where w is the weight of the node, n is the node, N is the set of all of the nodes.

Definition 2: Deviation is the distance between v and b, v is the input vector and b is the optimal match node, denoted as ε, the calculating method as the following formula (15):

$$\varepsilon = \sum_{i=1}^{d} (v_i - w_b)^2 \delta \quad (15)$$

Where d is the dimension of the input vector v, δ is the growth threshold.

Definition 3: the neighborhood of node n is the area that includes n and its sub-nodes.

The process of TGSOM is as following.

S1: Initialize

Firstly, all the vectors in input vector set will be standardized to an area of [0, 1], and their weight are assigned as the random values between 0 and 1. The growing threshold, denoted as δ, will be computed according to user's requirement.

S2: Training sample set

a. Select training sample vector v randomly from S, find the optimal match node b from the nodes of TGSOM, and then compute ε between b and v.

If $\varepsilon \leq \delta$, the weight of neighborhood of b will be adjusted as the following formula (16):

$$w_i(j+1) = \begin{cases} w_i(j), i \notin N_{j+1} \\ w_i(j) + \eta(j) \times (v_j - w_i(j)), i \notin N_{j+1} \end{cases} \quad (16)$$

$\eta(j)$ is the learning rate, when $j \to \infty$, then $\eta \to 0$; $w_i(j)$, $w_i(j+1)$ are the weight of node i before and after adjustment. N_{j+1} is the neighborhood of b in the (j+1)th training.

If $\varepsilon > \delta$, then generate a new sub node c, and w_c=v, and the adjust η, $\eta(j+1) = \eta(j) \times \alpha$, α is an adjustment factor of η, and $0 < \alpha < 1$.

b. repeat a until all the samples in v finish training.

S3: repeat S2 to the next training cycle, until there is no new node generated in the web.

S4: diminish η smoothly, and make fine rectify of weight of nodes, find b of v and adjust b's weight of neighborhood. In this step, there is no new node generated, and its purpose is to amend deviation. It is fit for the nodes that are generated in the later period of growing process.

In algorithm of TGSOM, whether or not generate new node depends on δ, but δ is uncertain, so it is not convenient to users, and can not determine the level of clustering. So we bring forward a spread factor SF, let $\delta = [d \times (1-SF)^2] / [1+1/n_t]$ (17) 0<SF<1, and when SF is small and δ is large, it could accomplish low level clustering, and when SF is large and δ is small, it could accomplish high level clustering. n_t is the total node number in the tth training.

Although the spread factor SF could be valued arbitrarily in the interval[0,1], but in order to obtain a rough cluster number in the first clustering, the initial spread factor SF is suit to choose a value in a lower interval [0,0.4]; after the text pre-processing, unsupervised feature selection, the first clustering, the latent semantic analysis(LSA) and some other series of operations, SF can be valued in slightly higher interval [0.4,0.7] to obtain the more accurate clustering. According to the need, if it would get a further more accurate clustering results, it can execute the 3rd clustering and SF = 0.8.

In this algorithm, different SF could be chosen in the same attribute set, or choose the same SF in different attribute sets to compare the result of clustering, and then select the optimal SF to clustering. As the establishment of the tree structure, the clustering number could be determined.

B. FKM Algorithm

FKM[6] algorithm is the combination of k-means and fussy theory, and fossilize the tough classification of k-means. We bring forward a modified subordination function u_{ij} [1], u_{ij} refers to the subordination degree of the ith sample to the jth category. We define a criterion function of clustering, that is $J = \sum_{j=1}^{k} \sum_{i=1}^{k} [u_{ij}]^{\beta} \| x_i - m_j \|^2$, and the constriction condition is $\sum_{i=1}^{k} \sum_{j=1}^{n} u_{ij} = c$. For every vector v_i, i = 1, …n, its subordination function u_{ij} is compute as the following formula (18):

$$u_{ij} = \frac{(1/\| v_i - x_j \|^2)^{1/(c-1)}}{\sum_{i=1}^{k} (1/\| v_i - x_i \|^2)^{1/(c-1)}}$$, j=1,2,…,k; $c > 1$ (18)

According to fussy theory, the FKM algorithm is described as following.

Input: clustering number K, parameter b and database with N objects

Output: K clusters and with a minimum sum of the square of deviation.

S1: get clustering number K and clustering center according to tree structure.

S2: in the process of the kth iteration, use u_{ij} to update the clustering center c_i of each category,

$$c_i = \frac{\sum_{j=1}^{n} (u_{ij})^c x_j}{\sum_{j=1}^{n} (u_{ij})^c}$$, i=1,…,k (19)

S3: for all the samples, the process will be finish if their subordination function won't change any more, if not, go back to S2.

V. Experiment and Analysis

In order to evaluate the model of TCFM and the algorithm of TGSOM-FS-FKM, we compare them with DSOM-FS-FCM 、 DSOM-FCM 、 DFKCN and FDMFC which based on FDM algorithm.

A. Evaluation of Text Clustering

Precision and Recall are often used as criteria of evaluation, and in this paper, we also use these two criteria. For each topic c and its corresponding clustering category,

Precision (C, c) =n1/(n1+n2) ;

Recall (C, c) =n1/(n1+n3);

F-measure= 2PR/(P+ R) (20)

P refers to Precision and R refers to Recall, n1 refers to the number of documents that belong to clustering C and their topic is c, n2 refers to the number of documents that belong to clustering C and their topic is not c, n3 refers to the number of documents that their topic is c but not classified to the clustering category C.

VI. Experiment and Result Analysis

In the experiment, we choose 260 documents as input, in them there are 50 respectively about the topic of economy, medicine and tourism, and there are 55 respectively about esthetics and computer. According to TCFM, firstly we get 8256 feature items in text process, and through DFFS, we get only 819, has a proportion of 9.92% to the whole feature items. Then execute TGSOM, after a clustering, and LSA, there are 630 feature items left, has a proportion of 7.63%, and go on executing TGSOM, after the second clustering, use an improved feature item selecting method of CHI to deal with the clustering result, and in this method, we configure different parameters and choose appropriate feature item numbers, the numbers are 536、439、364、293、229、164 respectively, and the proportion are 6.49%、5.32%、4.41%、3.55%、2.77%、1.99% respectively. In order to make compartment among DSOM-FS-FCM 、 DSOM-FCM 、 DFKCN and

FDMFC, we also use the improved CHI method and remain the same feature item numbers. We choose the optimal F-measure as the final result, and the result shows in Fig.4.

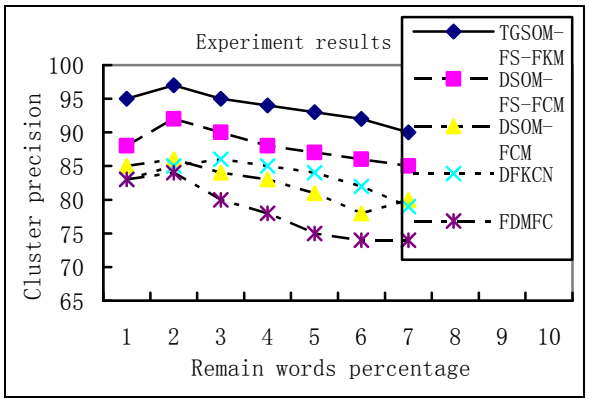

Figure 4. Experiment results

From the figure, it shows that the precision of DSOM-FS-FCM is less than TGSOM-FS-FKM, it is because there are mistakes of markup of text classification in DSOM, the F-measure of DSOM-FCM、DFKCN and FDMFC are not very high, because they didn't delete noisy words completely, and as a result affect their clustering precision. When their dimension is very high, FDMFC has some disturbance in the recognition of classification.

V. Conclusion

In this paper, we analyzed the characteristics of the text, which its number is huge and its storage structure is high-dimension and sparse, and also analyzed the shortage of current clustering methods. Based on above analysis, we proposed a new TCFM, and put forward a new text clustering method TGSOM-FS-FKM. In this method, it firstly makes preprocess of texts, and filter the majority of noisy words by using unsupervised feature selection method. Then it used TGSOM to execute the first clustering to get the rough classification of texts, and to get the initial clustering number and each text's category. And then introduced LSA theory to improve the precision of clustering and reduce the dimension of feature vector. After that it used TGSOM to execute the second clustering to get the more precise clustering result, and used supervised feature selection method to select feature items. Finally, it used FKM to cluster the result set. In the experiment, it remained the same number of feature items. By compared with DSOM-FS-FCM 、 DSOM-FCM 、 DFKCN and FDMFC, the result shows that TGSOM-FS-FKM algorithm would.

References

[1] YE Ping, Fuzzy K-means algorithms based on membership function improvement[J].Changchun Institute of Technology(Natural Sciences Edition),,2007,(01)

[2] HE Zhongshi,XU Zhejun,A New Methal Unsupervised Feature Selection for Text Mining[J]. Chongqing University(Natural Science Edition),2007(06)

[3] XIONG Zhongyang, ZHANG Pengzhao and ZHANG Yufang， Improved approach to CHI in feature extraction[J]. Computer Applications,2008,(02)

[4] GENG Xinqing,WANG Zhengou,DFKCN:A Dynamic Fuzzy Kohonen Neural Network and Its Application[J].Computer Engineering, 2003,(03)

[5] Wang Li Wang Zhcngou, TGSOM: A NEW DYNAMIC SELF-ORGANIZING MAPS FOR DATA CLUSTERING[J]. Electronics and Information Technology, 2003(03)

[6] GUO yan-fen; LI Tai,Design of fuzzy K-means-based fuzzy classifier[J]. Technical Acoustics,2007(04)

[7] GONG Jing, LI Ying-jie,Analysis and Comparison on Text Clustering Algorithm[J]. Hunan Environment-Biological Polytechnic,2006(03)

[8] R. Siegwart, I. R. Nourbakhsh, "Introduction to Autonomous Mobile Robots," Prentice Hall of India (pvt.) Ltd., 2005.

[9] G. Hamerly and C. Elkan, "Alternatives to the K-Means Algorithm That Find Better Clusterings," Proc. 11th Int'l Conf.Information and Knowledge Management (IKM '02), pp. 600-607,2002.

[10] P. Arabshahi, J. J. Choi, R. J. Marks 11, and T. P. Caudell, "Fuzzy control of backpropagation," in Proc. IEEE Int. Conf. Fuzzy System (FUZZ-IEEE '92), San Diego, CA, Mar. 1992.

[11] Z. Huang, M. Ng, H. Rong, and Z. Li, "Automated Variable Weighting in k-Means Type Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 5, pp. 657-668, May 2005.

[12] M. Laszlo and S. Mukherjee, "A Genetic Algorithm Using Hyper-Quadtrees for Low-Dimensional K-Means Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 4,pp. 533-543, Apr. 2006.

[13] D. Arthur and S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," Proc. 18th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA '07), pp. 1027-1035, 2007.

[14] Y. Cheung, "Maximum Weighted Likelihood via Rival Penalized EM for Density Mixture Clustering with Automatic Model Selection," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6,pp. 750-761, June 2005.

[15] D. Lieb, A. Lookingbill, and S. Thrun, "Adaptive Road Following using Self-Supervised Learning and Reverse Optical Flow," Stanford Artificial Intelligence Laboratory, Stanford University, 2005.

[16] Gath and A. Geve, "Unsupervised Optimal Fuzzy Clustering,"IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 11, no. 7,pp. 773-781, July 1989.

[17] S. Padhraic, "Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood," Statistics and Computing, vol. 10,pp. 63-72, 2000.

[18] Y. Cheung, "Maximum Weighted Likelihood via Rival Penalized EM for Density Mixture Clustering with Automatic Model Selection," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 6,pp. 750-761, June 2005.

Jinzhu Hu（1947-）, male，professor，Doctoral tutor，Research： Software Engineering and Distributed Information System.

Chunxiu Xiong（1984-）, female，Master，Research：Software Engineering and Distributed Information System.

Jiangbo Shu（1982-）, male，Ph.D.Research:Software Engineering, Chinese Information Processing.

Xing Zhou（1985-）, female，Master，Research：Software Engineering and Distributed Information System.

Jun Zhu（1984-）, male，Master，Research：Software Engineering and Distributed Information System.

