Concept Features Extraction and Text Clustering Analysis of Neural Networks Based on Cognitive Mechanism

Lin Wang[1], Minghu Jiang[1,2], Shasha Liao[2], Beixing Deng[3], Chengqing Zong[4], and Yinghua Lu[1]

[1] School of Electronic Eng., Beijing Univ. of Post and Telecom, Beijing, 100876, China
[2] Lab of Computational Linguistics, School of Humanities and Social Sciences, Tsinghua University, Beijing, 100084, China
[3] Dept. of Electronic Eng., Tsinghua University, Beijing, 100084, China
[4] State Key Lab of Pattern Recognition, Institute of Automation Chinese Academy of Science, Beijing, 100080, China
jiang.mh@tsinghua.edu.cn

Abstract. The feature selection is an important part in automatic classification. In this paper, we use the HowNet to extract the concept attributes, and propose CHI-MCOR method to build a feature set. This method not only selects the highly occurring words, but also selects the word whose occurrence frequency is middle or low occurring words that are important for text classification. The combined method is much better than any one of the weight methods. Then we use the Self-Organizing Map (SOM) to realize automatic text clustering. The experiment result shows that if we can extract the sememes properly, we can not only reduce the feature dimension but also improve the classification precise. SOM can be used in text clustering in large scales and the clustering results are good when the concept feature is selected.

1 Introduction

After a decade of emphasis on the study of brain mechanisms at the cellular molecular or genomic level, it is expected that future advances in brain science will promote the study of natural language processing (NLP). With the rapid development of the online information, automatic classification becomes one of the key techniques for handling and organizing the very large scale of text data. In the future, a fundamental breakthrough in text classification could be of benefit to diverse areas such as semantic nets, search engines, and natural language processing.

Text automatic classification based on cognitive science is a cutting-edge research topic both in studying brain cognitive systems and natural language processing. Extraction of brain cognitive principles improves understanding of natural language. Its theoretical models will lead to benefits both the cognitive science and the natural language processing. It will provide feedback to experimental methods concerning the validity of interpretations and suggestions, and enable us to create semantic methods which let the computer to understand language. Our aim is to understand the biological mechanisms of text classification and its role in perception and behavioural

D.-S. Huang, K. Li, and G.W. Irwin (Eds.): ICIC 2006, LNCS 4113, pp. 235 – 246, 2006.

simulation. Although neuroimaging methods by using localization of cognitive operations within the human brain can be applied to studies of neural networks, the conventional syntax techniques are still ineffective in natural language processing due to a lack of semantic understanding of relevance, in addition the concept attributes are much better to reflect the content of the documents, we can get a much better vector space by using the concept attributes and semantic information.

This paper is organized as follows: Section II presents the concept extraction method. Section III presents hierarchical clustering and SOM clustering. Section IV is about experiments and Section V summarizes the conclusions.

2 Concept Extractions

The experimental data is 68 words [1] are based on "Dictionary for Modern Chinese Syntax Information" and "Hownet" [2], which are described according to their syntax and semantic attributes, the feature set is consist of 50-dimension syntax features and 132-dimension semantic features. By using SOM neural network to train the 68 Chinese words including nouns, verbs and class-ambiguous words, we compare the fMRI experimental results of Li Ping et al [1] with the map results of neural networks for the three kinds of Chinese words. The neuroimaging localization of LiPing's brain experiments shows that there are obvious the overlapping of brain mapped distributing areas for the three kinds of words. In our SOM experiment [3], when we strengthen the role of syntax features, and weaken the role of semantic features, the overlapping of the mapped distributing areas for the three kinds of words can disappear. Whereas, when we weaken the role of syntax features, and strengthen the role of semantic features, the overlapping of the mapped distributing areas for the three kinds of words is increased. When we adpoted only semantic features to desciribe the three kinds of words, the distributing areas of mapped results are almost entirely overlapped. The experimental result shows that feature description plays an important role in the map area of the three kinds of words. In fact the response of human brain to Chinese lexical information is based mainly on conceptual and semantic attributes, in our accustomed conversation we pay seldom attention to Chinese syntax and grammar features, which is coincident with our experimental results, is also coincident with LiPing's.

We extract the concept attribute from the word as the reflection of the text, which will describe the internal concept information, and get the relationship among the words. The information of the concept extraction comes from HowNet [2] and the synonymy dictionary, use the DEF term of the Chinese word, which descript the word with defined concept attribute, in order to construct the feature reflection of the documents.

2.1 Analysis of the Feature Set

When we extract the concept attributes to form the feature set, we convert a lot of words into the concept features, and get rid of the influence of the synonymy and dependence, which makes the classification precise much higher [4]. However, because of the mass of weak concept and the words which are not in the HowNet, some Chinese words are given a comparatively lower weight and become the middle

or low occurring feature. In addition there are still some specialty words and proprietary words which are only occur in one category and are not highly occurred in the whole documents and are very important for classification, both of these words need a strategy to get a higher weight and contribute more in text classification, thus we analysis and experiment on the weighting methods in the following parts.

2.2 CHI Selection Method

The CHI (χ statistics) weight method's formula is shown as follows [4]:

$$\chi^2(t,c) = \frac{N*(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} . \tag{1}$$

$$\mathrm{x}^2_{\max}(t) = \max_{i=1}^{m} \chi^2(t,c_i) . \tag{2}$$

Here, N is the total document number of the training set, c is a certain category, t is a certain feature, A is the number of the document which belong to c and t, B is those which do not belong to c but contain t, C is those which belong to c but do not contain t, D is those which do not belong to c and do not contain t.

CHI method is based on such hypothesis: if the feature is highly occurred in a specified category or highly occurred in other categories, it is useful in classification. Because CHI take the occurrence frequency into account, it prefers to select highly occurred words, and ignore the middle and low occurred words which maybe important in classification.

2.3 MCOR Selection Method

The MCOR (Multi-Class Odds Ratio) weight method's formula is shown as follows [4]:

$$MCOR(t) = \sum_{i=1}^{m} p(C_i) \left| \log \frac{P(t/C_i)(1-P(t/C_{else}))}{(1-P(t/C_i))P(t/C_{else})} \right| . \tag{3}$$

Here, P(Ci) is the occurrence probability of category Ci, P(t / Ci) is the occurrence probability of the feature t when category C is occurred, P(t / Celse) is the occurrence probability of the feature t when category C is not occurred. When P(t / Ci) is higher or P(t / Celse) is lower, the weight of MCOR is higher. Therefore, the MCOR selects the features which are mainly occurred in one category and nearly not occurred in other categories. Because it does not consider the occurrence frequency of the features, it prefers to select the words which are middle or low occurred in the document while highly occurred words are always occurred in more than one categories.

2.5 The Comparing Result of Seven Weighing Methods

We select seven common weighing methods of the features and test their performance, and focus mainly on their selection strategy and classification precision. The experimental result is shown in Fig. 1.

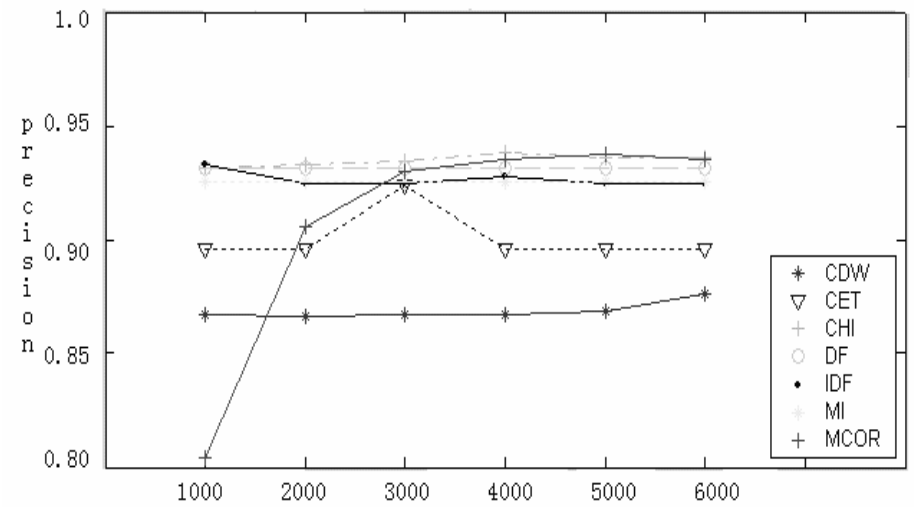

Fig. 1. The average of seven different weighing methods. Y axis is the average precision, and x axis is the feature dimension of the training set.

From the analysis of the selected features, we find that:

1. The DF (Document Frequency), TF-IDF (Term Frequency-Inverse Document Frequency), CET (an improved method of Information Gain), CDW (Category-Discriminating Word) and CHI methods prefer the high-occurred words and they are greatly related. In our experiment, CHI is the best method.
2. The MCOR method mainly chooses the middle and low occurred features, so its classification precision is low when the reduction rate is high. But with the increase of the feature dimension, its precision is increased highly and when the feature dimension is above 4000, its precision is higher than CDW, CET, DF, TF-IDF and MI (Mutual Information) methods.
3. The MI method mainly selects the high and middle occurred feature, it can get a good classification precision but with the increase of the feature dimension, the precision is not improved visibly.

2.6 Combined Method of CHI-MCOR

Because the MCOR mainly selects the words whose occurrence frequencies are middle or low, its classification precise is low when the reduction of feature dimensions is high. But with the increase of feature dimensions, its precise is improved to an appreciable level. The CHI prefers to select the words whose occurrence frequencies are high, and it is one of the best feature selection methods. As a result, when we combine these two methods, we can make the advantages together and get a high classification precise. Therefore, we give a combined weight method based on the CHI and MCOR:

$$V(t) = \lambda V_{CHI}(t) + (1-\lambda)V_{MCOR}(t), \qquad 0 < \lambda < 1. \tag{4}$$

Where, V_{CHI} is the weight of feature t for the CHI method, V_{MCOR} is the weight of feature t for the MCOR method. When we analysis the weights given by these both methods, we find that the average weight of the features are different. For example, when the reduction of feature dimensions is 50%, the range of the CHI weights is (2.1, 6.81), while that of the MCOR weights is (1.15, 1.76). Because the CHI gives a much higher weight to all the features and its swing is wider, we should give a comparatively lower value to λ, else the value depends too much on the CHI and the combined weight method is meaningless. Therefore we need a proper value of λ. According to experience, let us assume that when the average weight of the CHI and MCOR are the same, we can get both advantages of the two and the classification precise will be the highest, therefore the best λ cab be evaluated as follows:

$$\frac{\lambda}{1-\lambda} = \frac{Mean(MCOR)}{Mean(CHI)}. \tag{5}$$

From Fig. 2 we can see that the combined weighing method is much better in politics category, it means that there are a lot of important words in politics category which are not highly occurred. Therefore, when the combined CHI-MCOR method is used, its precision is 3.66% higher than the CHI method. In fact, we find that the top ten occurred words in politics category which are not very high in the total statistics.

In order to analysis the best value of λ, we vary λ from 0 to 1.0. From the experiment, we found that when λ is 0.3, the classification precise is the highest. This result accords to our hypothesis. Meanwhile, we find that when we use the combined weight method, the precise is always higher than other methods. For example, when λ is 0 or 1, it is the precise of the MCOR method or the CHI method. When λ is 0.3, the precise is 94.0359%, which is 0.61% higher than the CHI's, 1.074% higher than the MCOR's. When we use the combined CHI-MCOR method, its precise is 3.66% higher than we only use the CHI method. In fact, when we

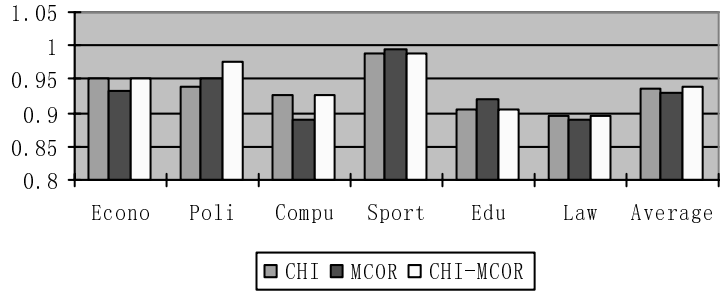

Fig. 2. The precision of the six categories in three weighing methods, the CHI, MCOR and CHI-MCOR. Y axis is the classification precision, and x axis is the test categories, the last one is the average precision.

statistic the top ten of the occurred words in politics category, we find that they are not very high in the total statistics. Because there are some words which are not highly occurred but useful and important for text classification, we use the combined CHI-MCOR method to take balance in the high occurring ones and the middle occurring ones. This method not only selects the highly occurring words, but also selects the word whose occurrence frequency is middle or low, and its features only belong to one or two categories. The experimental result shows that the combined method is much better than any one of the weight methods.

3 Hierarchical Clustering and SOM Clustering

3.1 Hierarchical Clustering

Hierarchical clustering creates a cluster tree to investigate grouping in input data, simultaneously over a variety of scales of distance. The result of hierarchical clustering can be graphically represented by a multi-level hierarchy (dendrogram), where clusters at one level are joined as clusters at the next higher level. The root is the whole input data set, the leaves are the individual elements of input data, and the internal nodes are defined as the union of their children [5]. Each level of the tree represents a partition of the input data into several groups (clusters). We can investigate different scales of grouping in text data, this allows us to decide what scale or level of clustering is most appropriate in our application.

3.2 Self-Organizing Map (SOM) Clustering

The SOM is based on research of physiology and brain science which is proposed by Kohonen [6]. By using self-organized learning to the network enables the similar neuron in function to be nearer, the different neuron in function to be more separate. During the learning process, no predefined classes of input data are clustered automatically and enable the weight distribution to be similar to input's probability density distribution. The SOM learns to recognize groups of similar input vectors in such a way that the neurons physically near each other in the output layer respond to the similar input vectors, i.e., the lesser the distance, the greater the degree of similarity and the higher the likelihood of emerging as the winner. The SOM can learn to detect regularities and correlations of input data, its training is based on two principles [6]:

1). *Competitive learning*: the prototype vector most similar to an input vector is updated so that it is even more similar to it.
2). *Cooperative learning*: not only the most similar prototype vector, but also its neighbors on the map are moved towards the input vector.

The SOM not only can adapt the winner node, but also some other neighborhood nodes of the winner are adapted, it can learn topology and represent roughly equal distributive regions of the input space, and similar inputs are mapped to neighboring neurons. The SOM consists of input layer and output layer, which is constructed by competitive learning algorithm based on above two principles. Unlike other cluster

methods, the SOM has not distinct cluster boundaries, therefore, it requires some background knowledge to solve it [7]. Here we adopt the best Davies-Bouldin index to classify cluster boundaries. The choice of the best cluster can be determined by the Davies-Bouldin index [5]. It is a function of the ratio of the sum for within-cluster distance and between-cluster distance. Optimal clustering is determined by [5]:

$$V_{DB} = \frac{1}{N}\sum_{k=1}^{N} \max_{k \neq l} \frac{S_N(D_k) + S(D_l)}{T_N(D_k, D_l)} \tag{6}$$

Where N is the number of clusters, D is a matrix of the data set X, S_N *is* the within-cluster distance between the points in a cluster and the centroids for that cluster and T_N *is* the between-cluster distance from the centroid of one cluster to the other. The optimal number of clusters is the one that minimizes V_{DB}. If the clusters are well separated, then V_{DB} should decrease monotonically over time as the number of clusters increases until the clustering reaches convergence.

3.3 Supervisory Initiative Learning Based on Cognitive Mechanism

A man learns and acquires knowledge by the mode of "learn - practise - relearn - repractise", it suggests us that neural networks are trained by supervisory initiative learning, we introduce the feedback learning to text classification, thus the classification system is extended into a mode of "training - classification - feedback retraining". During the training process, the worth classification results can be selected, it can update the parameters of classification model according to feedback results, and reflect the cerebral cognitive process. We adopt Counterpropagation Networks which is proposed by Robert Hecht-Nielson, it consists of input layer, Kohonen layer and Grossberg output layer. By combining the unsupervisory training and supervisory training to form the network of separate structure, the network structure is closer to brain's one, the semantic concept features of text are preprocessed as input vectors, the unsupervisory training is finished in Kohonen layer, the classification information is extracted by the comparability of input data. The supervisory training is realized in Grossberg layer, the weights are updated by the difference between the expected output and the real output, and realize the category expression. The supervisory training can provide efficient codes and its classifiable boundary approaches a Bayes' boundary.

4 Experiments

According to above analysis, we extract the concept attributes by the combined CHI-MCOR method to build a concept feature set of 500 dimensions, is consist of both sememes and words, then we use the SOM network to realize automatic text clustering. The experimental corpus comes from the People Daily from 1996 to 1998. The corpus is unbalanced, and the training set includes 1205 texts (250 economy texts, 175 politics texts, 130 computer texts, 300 sport texts, 150 education texts, 200 law texts). The test set is another 755 texts of above 6 classes.

4.1 Hierarchical Clustering Experiment

The concept feature sets is used in clustering experiments. Fig. 3 shows that the clustering results of concept features has obvious cluster groups, form several wave crests and hiberarchy, there are obvious distances among different groups in the 1205 training texts.

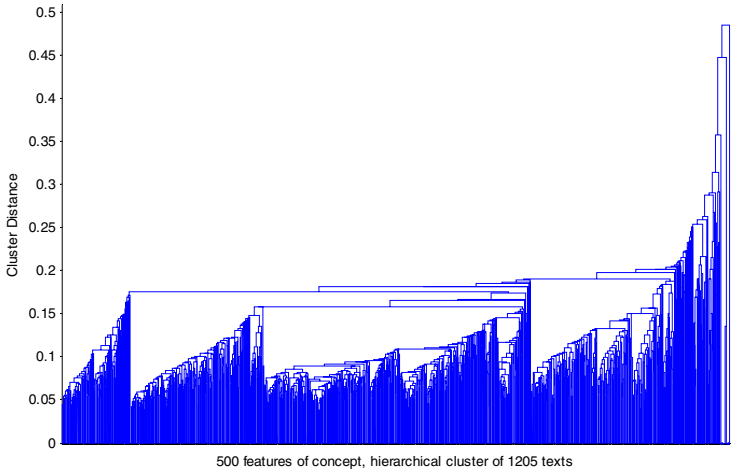

Fig. 3. Hierarchical clustering for 500 features of concept and 1205 texts

4.2 SOM Clustering Experiment

The SOM's initialization is linear with small random initial weights, and batch training algorithm is used in two phases of rough training and fine-tuning. The size of SOM output layer is 15×11 nodes which depends on dimension number and distribution of input features, the training time is 4+11 seconds which run in a P4 180M computer. Fig. 4 shows the U-matrix (left figure) and D-matrix (right figure) of the SOM clustering by using the 1205 texts of 500 concept features. The 'U-matrix' shows distances between neighboring units and thus visualizes the cluster structure of the map, it has much more hexagons in the visual output planes because each hexagon shows distances between map units. While D-matrix only shows the distance values at the SOM map units. Clusters on the U-matrix are typically uniform areas of low values (white) which mean small distance between neighboring map units, and high values (black) mean large distance between neighboring map units and thus indicate clustering borders. There are more clustering borders of high values (black) in Fig. 4, it shows that there are more small clusters for texts of concept features, several white zones (uniform areas of low values) are encircled by black or gray clustering borders. It shows same as hierarchical clustering that the between-cluster distance of concept features is obvious.

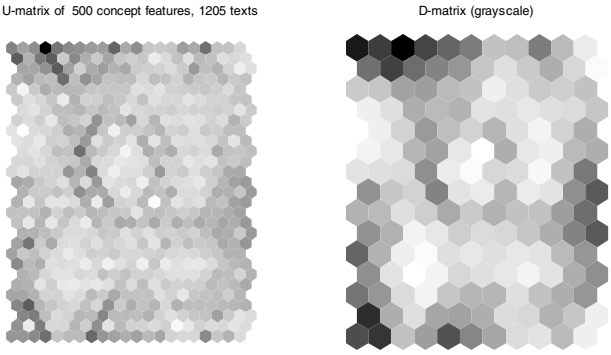

Fig. 4. U-matrix (left figure) and D-matrix (right figure) of the SOM, the training data are 1205 texts of 500 concept features, the SOM's initialization is linear, and batch training algorithm is used in two phases of rough training and fine-tuning

Because the SOM has not distinct clustering boundaries, in order to find and show the borders of the SOM clusters, we use the k-means clustering to find an initial partitioning [8], the experimental results show that the values of important variables change very rabidly. We can assign colors to the map units such that similar map units correspond to similar colors. Fig. 5 is the SOM clustering results by using the 1205 texts of 500 concept features, the left figures show the Davies-Boulding (DB) clustering index [5]; and the right figures show the SOM clustering by color code which is minimized with best clustering. According to DB index, we can find that VDB is almost monotonously decreased with increase of clustering groups. The number of the best clusters is 14 (corresponding to their minimum VDB values) for concept features.

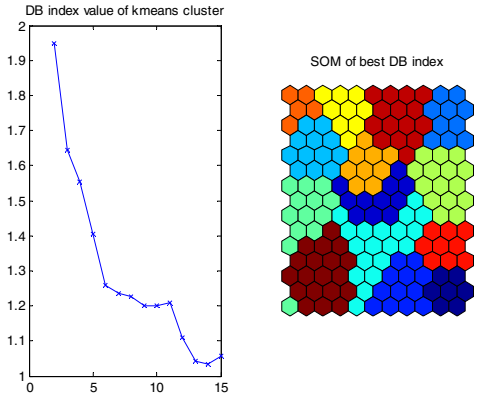

Fig. 5. Davies-Boulding clustering index (left figure), and the SOM clustering by color code which is minimized with the best clustering (right figure). Training data are 1205 texts of 500 concept features.

The left figure in Fig. 6 is the number of map samples in each unit; it shows the distribution of the input data on the output map plane. Because the significance of the components with respect to the clustering is harder to visualize, therefore we adopt distance matrix with color codes which is minimized with the best clustering on the right figures of Fig. 6. Small hexagons indicate clustering borders (corresponds to large distance between neighboring map units on U-matrix), it shows more small clusters for texts of concept features.

By comparison of the hierarchical clustering and the SOM clustering, the results show distinctly easily that between-cluster distance of the texts of concept features is obvious. Both results of the SOM clustering and artificial classification have a good corresponding relationship in the rough. A group or several groups in SOM clustering may correspond to some a class of artificial classification. There exist some fuzzy output nodes (hexagon) in Fig. 6, i.e., there are different artificial classes in same output nodes or same color area. The clustering qualities of SOM are evaluated by the precision P, recall R and parameter $F1$. The formulas of precision and recall for class k are defined as follows:

$$\Pr ecision_k = \frac{AcorrectNum_k}{AtotalNum_k} \tag{7}$$

$$\mathrm{Re}\, call_k = \frac{CorrectNum_k}{TotalNum_k} \tag{8}$$

Where, the $AcorrectNum_k$ is the number of the documents of the class k which are correctly judged by a computer; $AtotalNum_k$ is the number of the documents of the class k which are judged by a computer. The $CorrectNum_k$ is the number of the documents of the class k which are correctly classified; $TotalNum_k$ is the number of the documents of the class k in a standard solution.

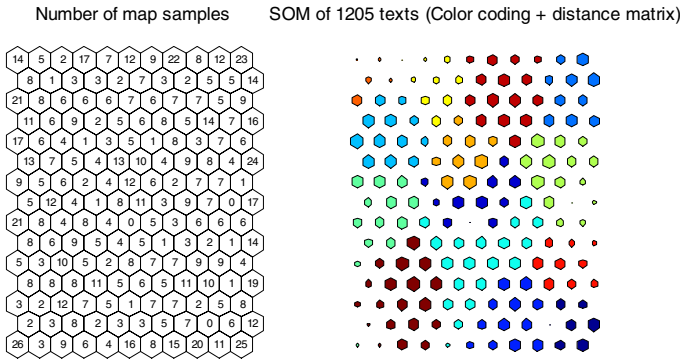

Fig. 6. The distribution of the input data on the map (left figure), the digital in each hexagon is the number of map texts; distance matrix with color codes is shown on the right figure, small hexagons on the D-matrix indicate clustering borders (corresponds to large distance between neighboring map units on U-matrix). Training data are 1205 texts of 500 concept features.

Then the average values of precision, recall and $F1$ can be obtained as the clustering results of SOM:

$$\mathrm{P} = \frac{\sum_{k=1}^{m} \mathrm{Pr}\,ecision_k}{m} \quad R = \frac{\sum_{k=1}^{m} \mathrm{Re}\,call_k}{m} \quad F1 = \frac{2\mathrm{P} \times R}{P + R} \tag{9}$$

Table 1 shows that the clustering performance of concept features. The experimental result shows that the SOM can be used in text clustering in large scales and the clustering results are good when the concept feature is selected. If we can extract the sememes properly, we can not only reduce the feature dimension but also improve the classification precise.

Table 1. SOM Clustering Results of Concept Feature sets

	Economy	Polity	Computer	Sport	Education	Law	Average	$F1$
P	96.5	98.0	92.8	99.1	93.1	91.6	95.18	93.16
R	94.2	85.3	95.2	94.6	89.8	88.2	91.22	

5 Conclusions

Because there are some words which are not highly occurred but useful in text classification, while the words with high occurrence frequency is usually useful, except the words in the stop word dictionary which are frequently used in the text but useless in classification, our CHI-MCOR method to take balance in the high occurring ones and the middle occurring ones. This method not only selects the highly occurring words, but also selects the words whose occurrence frequencies are middle or low but only occur in one or two categories. It is much better than CHI or MCOR method alone. When we use concept as the feature of text classification, we can efficiently reduce the feature dimension and reflect the original feature space to a more stable one. The experimental result shows that the SOM can be used in text clustering in large scales and the clustering results are good when the concept feature is selected. Between-cluster distance of the texts of concept features is bigger than that of texts of word features.

References

1. Li, P., Jin, Z., Tan, L. H.: Neural Representations of Nouns and Verbs in Chinese: an fMRI Study. NeuroImage, 21 (2004) 1533-1541
2. http://www.keenage.com
3. Jiang, M., Cai, H., Zhang, B.: Self-organizing Map Analysis Consistent with Neuroimaging for Chinese Noun, Verb and Class-ambiguous Word. Advances in Neural Networks – ISNN2005: Lecture Notes in Computer Science, Springer-Verlag Heidelberg, 3498 (2005) 971-976

4. Liao, S., Jiang, M.: An Improved Method of Feature Selection Based on Concept Attributes in Text Classification. Advances in Natural Computation, Lecture Notes in Computer Science, Springer-Verlag Heidelberg, 3610 (2005) 1140-1149
5. Davies, D., Bouldin, D.: A Cluster Separation Measure. IEEE Transactions on Pattern Analysis and Machine Intelligence - I, 2 (1979) 224-227
6. Kohonen, T.: The Self-organnized Map. Proceedings of the IEEE, 78 (1990) 1464-1480
7. Vesanto, J., Alhoniemi, J.: Clustering of the Self-organizing Map, IEEE Transactions on Neural Networks, 11(3) (2000) 586-600
8. Wang, L., Jiang, M., Lu, Y. et al: Self-organizing Map Clustering Analysis for Molecular Data. ISNN06, Lecture Notes in Computer Science, 3971 (2006) 1250-1255