A Feature Selection Methods Based on Concept Extraction and SOM Text Clustering Analysis

Lin Wang[†], *Minghu Jiang*[††], *Shasha Liao*[††], *Yinghua Lu*[†],

[†]School of Electronics & Engineering, Beijing Univ. of Post and Telecom., Beijing, 100876, China
[††]Lab of Computational Linguistics, School of Humanities and Social Sciences, Tsinghua Univ., Beijing, 100084

Summary

The feature selection is an important part in automatic classification. In this paper, we use the HowNet to extract the concept attributes from word to build a feature set. However, as the concept defi4nition sometimes is too weak in expression, we set a shielded level in the sememe Tree and filter the concept attributes which can not give enough information for classification, and reserve the word whose definition is too weak in expression. By this method, we build a feature set composing of both sememes from the HowNet and the Chinese words. We also give different sememes different values according to their expression ability and relation to the word when we extract them from the word. After comparing the weight theories and classification precise, we give the CHI-MCOR weight method, which is derived from two normal 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. The combined weight method makes a good balance between the fuzzy words which have a high occurrence and the dividing words which have a middle or low occurrence, and the classification precise is higher than other weight methods. 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, word features data nevertheless exhibit some clusters.

Key words:
Concept Attributes, Self-Organizing Map, Clustering, Text Classification.

Introduction

Automatic classification of Chinese text is a process which classifies the Chinese documents to several categories by their content. With the rapid development of the online information, automatic classification becomes one of the key techniques for handling and organizing the text data. It always has two main parts: the feature selection which reflects the documents to a feature vector space and weights the components of the vector; the classifier, which classifies the documents to the right category by their feature vectors. Because the huge data of the document set and the hardness of reflecting the documents to feature vector, we need to construct a proper feature set [1]. Nowadays, the Chinese feature selection methods are mainly based on the word feature and concept attribute. The word feature is the most popular one, because it is easy to be understood and handled. However, as the word is sometimes synonymous and dependent on its context, the word feature always ignores the relationship among the words and isolates all the words which are semantically related [3].

Because the concept space is much smaller than the word one, and the components are comparatively independent, 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 [2], so that we choose concept attribute as the main component of our feature selection method [3].

Feature weight method is also important in feature selection because it cannot only give a proper threshold to reduce the feature dimension to a computable one but also strengthen the important features. There are two kinds of feature selection method: the first one is based on threshold filtering, including DF (Document Frequency), TF-IDF (Term Frequency–Inverse Document Frequency), MI (Mutual Information), CHI(χ statistics), IG (Information gain) and so on [4][6]. After word features and concept features are obtained, we can use to clustering algorithms to realize text classification. Clustering analysis is a way to examine similarities and dissimilarities of observations. Data often fall naturally into groups or clusters, similar inputs should be classified as the same cluster and dissimilar inputs should be classified as the different clusters. Clustering is realized by unsupervised learning, no predefined classes, those within each cluster are more closely related to one another than objects assigned to different clusters. The data are classed into subgroups or clusters, such that the distance of data items within the same cluster (intra-cluster variance) is small and the distance of data items stemming from different clusters (inter-cluster variance) is large [7]. The clustering analysis is more appropriate for some aspects of biological learning, human and social sciences and related areas. The goal of clustering analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered.

Manuscript received January,25 2006.
Manuscript revised Feb., 2006.

This paper is organized as follows: Section II presents the concept extraction method. Section III describes the combined feature weight method, compares and analysis the virtues and shortcomings of these methods. Section IV presents hierarchical clustering and SOM clustering. Section V is about experiments and Section VI summarizes the conclusions.

2. The Concept Extraction Method

The information of the concept extraction in this paper comes from HowNet [5], which is not only a semantic dictionary, but also a knowledge system referring to the concept of Chinese words and the relationship among them. So we use the DEF term of the Chinese word, which descript the word with defined concept attribute, to construct the feature reflection of the documents. In our method, 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. Because there are 24,000 words in the HowNet and only nearly 1,500 concept attributes, the feature will be reduced to a stable dimensionality space with little information lose.

2.1 **Concept Weight Method Based on HowNet**

Because the different concept attributes have different expression abilities, it is unwise to give every attribute the same weight. In fact, we need a strategy to give different expression abilities to these attributes, so we consider two factors to weight them, one is the height of the weight node in a concept tree; the other is the number of the child nodes of the weight node. The height of the node is the most important factor because it shows the detail degree of the concept. Also, when a node have more child nodes, it means that in the cognize world, this concept is more complex and has more detail concept, and people would use its child nodes more and treat this concept as a more abstract one, so it should have a comparatively lower weight. Moreover, because the nine concept trees in HowNet are not equal, we give a different root weight to treat them differently. A weighting formula is shown as follow:

$$W_{ik} = Wtree_i \cdot [\log((Droot_{ik} + 1)/2) + \mathrm{a} + \frac{1}{Deep_k + \mathrm{b}}] \, .$$

Where，W_{ik} is the weighting of node k in tree i; $Wtree_i$ is the weighting of the tree root i, in case that there are nine trees in Hownet and they contribute differently in classification, and we give different weight to different trees; $Droot_{ik}$ is the distance between node k and the tree root i; $Deep_k$ is the number of the child nodes of node k; a and b are the tempering factors, which are used to control the weighting range. According to the experiments, we set a=1 and b=5.

2.2 The Abstract Concept Attributes and the Shielded Level in the Content-Tree

If the DEF term of a certain word contains only abstract concept, which has a weak expression ability, it means that this DEF term does not describes the word precisely and the information gain is not enough. So we cannot extract all the words into concept attributes. Here, we give a strategy to make a balance between the original words and the extracted concept attributes.

We use the concept tree of HowNet to calculate the expression ability. By a selected shielded level, we divided these nodes into two parts, the strong ones and the weak ones. Because we mainly use the level of the node to decide its expression ability, we set a level threshold, which is called shielded level. For a word, if all the attributes in the DEF are above the shielded level, we consider that these attributes are weak in expression and give less information than the original word, and we do not extract this word into concept. The formula calculating the concept expression ability of a word is shown as follow:

$$\mathrm{f}(c) = \max_{j=0}^{m} \; k(c_j) \, .$$

Where, $k(c_j)$ is the weight of attribute i in the DEF term of word c; m is the number of attributes in DEF term. This formula calculates all the attributes in a DEF term and decides whether the attribute or the word should be added to the feature set. If there is at least one attributes whose levels are higher than the shielded level, the expression ability of the DEF terms are enough and we added them into the feature set. Otherwise, the original word is added.

3． **Combined Weight Method**

3.1 The Analysis of the Feature Set

When we extract the concept attribute 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. 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. And 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 in classification. Both of these words need a strategy to get a higher weight and contribute more in text classification.

3.2 The Comparing Result of Seven Weight Methods

We select seven common weight methods and test them, and focus mainly on their selection strategy and classification precise. The experimental results are given in Fig. 1.

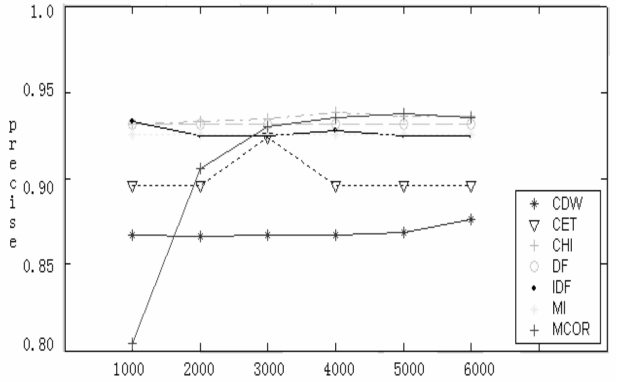

Fig. 1 Seven different weight methods, *y* axis shows the average precise, and *x* axis shows the feature dimension of the training set.

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

(i) The DF, TF-IDF, CET (an improved method of IG), CDW and CHI methods prefer the high occurred words and they are highly related. In our experiment, CHI is the best method in our experiments, which accords with the research of Yang [6]

(ii) MCOR method mainly chooses the middle and low occurred feature, so its classification precise is low when the dimension reduction rate is high. But with the increase of the feature dimension, its precise is increased highly and when the feature dimension is above 4000, its precise is higher than CDW , CET , DF , TF-IDF and MI methods.

(iii) MI method mainly selects the high and middle occurred feature, it can get a good classification precise but with the increase of the feature dimension, the precise is not improved visibly.

3.3 CHI Weight Method

The CHI weight method's formula is as follow:

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

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

Where, *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 category *c* and contain feature *t*, *B* is those which do not belong to category *c* but contain feature *t*, *C* is those which belong to category *c* but do not contain feature *t*, *D* is those which do not belong to category *c* and do not contain feature *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 ignored the middle and low occurred words which maybe important in classification.

3.4 MCOR Weight Method

The MCOR weight method is calculated as follow [1]:

$$MC-OR(t) = \sum_{i=1}^{m} P(C_i) \,|\log \frac{P(t/C_i)(1-P(t/C_{else}))}{(1-P(t/C_i)P(t/C_{else}))}| .$$

In this formula, *P*(*C_i*) is the occurrence probability of category *C*, *P*(*t*/*C*_i) is the occurrence probability of the feature *t* when category C_i is occurred, $P(t/C_{else})$ is the occurrence probability of the feature *t* when category *C* is not occurred. When $P(t/C_\mathrm{i})$ is higher or $P(t/C_{else})$ is lower, the weight of MCOR is higher. So, 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 feature, it prefer 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.

3.5 Combined Weight Method

Because MCOR mainly selects the words whose occurrence frequencies are middle or low, its classification precise is low when the dimension reduction is high. But with the increase of feature dimension, its precise is improved to an appreciable level. And CHI prefers to select the words whose occurrence frequencies are high, and it is one of the best feature selection methods [6]. As a result, when we combine the two methods, we can make the advantages together and get a high classification precise [7]. So, we give a combined weight method based on CHI and MCOR:

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

V_{CHI} is the weight of feature *t* of the CHI method, V_{MCOR} is the weight of feature *t* of the MCOR method. When we analysis the weigh given by the two methods, we find that the average weight of the features are different. For example, when the dimension reduction is 50%, the range of the weight of CHI is (2.1, 6.81), while the range of the weight of MCOR is (1.15, 1.76). Because CHI gives a much higher weight to all the features and its swing is wider, we should give a comparatively lower value to λ. If not, the value depends too much on CHI and the combined weigh method is meaningless. So we need a proper value of λ . According to our experience, we suppose that when the average weight of CHI and MCOR are the same, we can both get the advantage of the two and the classification precise will be the highest. So we think the best λ is as follow:

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

3.6 Relative Computation of Concepts and Features Extraction

Because some signs are used to describe semantic concepts in HowNet, so the correlation need be calculated between sememes and its concept in DEF term. The weight of sememecan be calculated by the following formula:

$$v(c_j) = freq(c) * k(c_j) * relate(c_j, c) .$$

Where, $v(c_j)$ is the feature value which includes sememe information, $freq(c)$ is a frequency of concept c in the text, $k(c_j)$ is the weight of attribute i in the DEF term of word c, $relate(c_j,c)$ is a relativity between sememes c_j and concept c, the relativity of independent sememes is 1.0, the relativity of sememe descriptions is 0.7, the relativity of sign sememes is among (0, 1) which depends on the different sign.

4. Hierarchical Clustering and SOM Clustering

4.1 Hierarchical Clustering

Hierarchical clustering creates a cluster tree to investigate grouping in input data, simultaneously over a variety of scales of distance [9]. 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 [9]. Each level of the tree represents a partition of the input data into several groups (clusters). We can investigate different scales of grouping in molecular data, this allows us to decide what scale or level of clustering is most appropriate in our application.

4.2 The Self-Organizing Map (SOM)

SOM is based on research of physiology and brain science which is proposed by Kohonen [10]. By using self-organized learning the network enables the similar nerve cell in function to be nearer, the different nerve cell in function to be more separate. During learning process, no predefined classes of input data are sorted automatically and enable the weight distribution to be similar to input's probability density distribution. SOM learns to recognize groups of similar input vectors in such a way that neurons physically near each other in the output layer respond to 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. SOM can learn to detect regularities and correlations of input data, its training is based on two principles [8]:

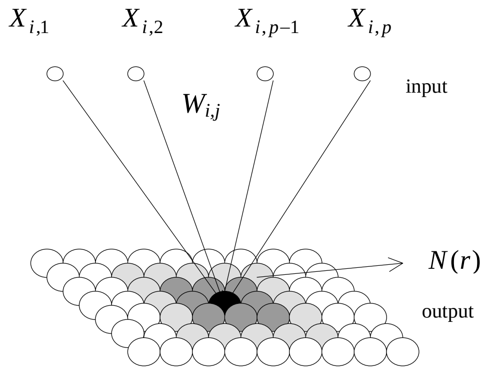

Fig. 2 SOM Neural Network

(i) Competitive learning: the prototype vector most similar to an input vector is updated so that it is even more similar to it.

(ii) Cooperative learning: not only the most similar prototype vector, but also its neighbors on the map are moved towards the input vector.

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. SOM consists of input layer and output layer, which is constructed by competitive learning algorithm. Each neuron in input layer is linked by the weight $W_{i\ j}$ to each neuron of output layer, the neuron within the area $N(r)$ around the winner neuron r in output layer obtain excitement in different degree, the neurons besides $N(r)$ are restrained. The area of $N(r)$ decreases monotonically over iteration number t, in finally there is only the remains of one neuron, it reflects the attribute of a kind of samples. SOM learning process is shown as follows [11, 12]:

When iteration number t=0, input data sample of no predefined classes $X = \{X_i \in \Re^p : i = 1,2,\Lambda\ ,n\}$, initial weight is put: $\{W_{i,j}, i, j=1,2,\cdots ,m\}$. When t<$T_{\max}$, randomly select $X_i(t)$ in X set:

Find out $r = \arg\min_s \{\|X_i(t) - W_s(t)\|\}$.

Iteration $W_s(t+1) = W_s(t) + \alpha_t \cdot e^{-dist(r,s)^2/\sigma_t^2} [X_i(t) - W_s(t)]$,

$$\forall s \in N_t(r)\,,$$

$$W_s(t+1) = W_s(t)\,, \qquad \forall s \notin N_t(r)\,.$$

Update t+=1, N_t=N_0-t(N_0-1)/T_{max}, α_t = α_0 (1-t/T_{max}), $\sigma_t = \sigma_0$ -t(σ_0 - σ_f)/T_{max}.

Here, m is output array size, $T_{\max}$ is the max iterative number, N_0 is initial neighbor threshold, α_0 is initial learning rate, σ_0 and σ_f are the control parameter of step length, $dist$(r, s) is a distance between neuron r and neuron s in the output array. N(r) and α_t decrease monotonically over iteration number t.

Unlike other cluster methods, the SOM has not distinct cluster boundaries, therefore, it requires some background knowledge to solve it. 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 [9]. It is a function of the ratio of the sum for within-cluster distance and between-cluster distance. Optimal clustering is determined by [9]:

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

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.

V. Experiments

This system is run in Windows XP, and the coding tools are VS.Net and Matlab7.0. The corpus comes from the People Daily from 1996 to 1998. The corpus is unbalanced, and the training set includes 1205 texts, here 250 texts belongs to economy, 175 texts belongs to politics, 130 texts belongs to computer, 300 texts belongs to sport, 150 texts belongs to education, 200 texts belongs to law. The test set includes 755 texts of above 6 classes.

5.1 Experimental Result of the Concept Extraction with Shielded Level

Fig.3 shows that only uses original words or concept attributes are both not very suitable, if we only use the concept attributes without any shielded levels, the precise is 90.9%, which is the lowest. And when we choose a proper level, for example, level 6, the precise is 93.7%, which is the highest. Moreover, when we use concept attribute as the feature, the difference among different categories are less than that when we use word features. This is probably because the feature selection based on original words depends much on the categories because if there are more special words in this field, it is easier to classify it from others. But when we use concept attributes, this difference between categories seems to be smaller and the curve seems to be much smooth.

The experimental result shows that concept extraction method can efficiently reduce the feature dimension, in feature dimension reduction, we do not lose useful information and the classification precise is much better because it filters the unnecessary noises.

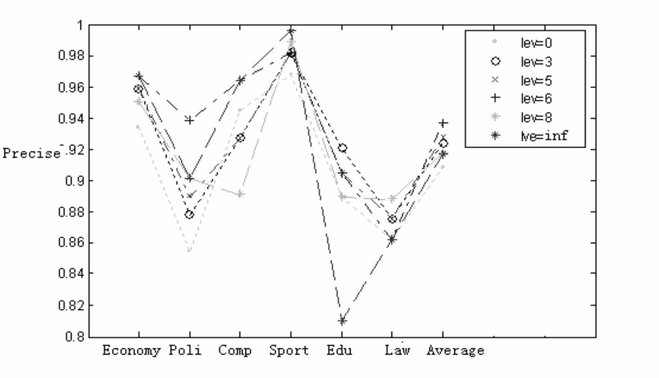

Fig. 3 This is the classification precise of the system with different shielded levels, The y axis is the classification precise, and x axis is the categories of the classification and the last one is the average precise as the precise of the system.

5.2 Experimental Result of the CHI-MCOR

In order to analysis the best λ value, λ is varied from 0 to 1.0. Fig. 4 shows that when λ is 0.3, the classification precise is the highest, when we use the combined weight method, and 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 CHI method. In our experiment, when λ is 0.3, the precise is 94.0359%, which is 0.61% higher than CHI, 1.074% higher than MCOR.

Fig. 5 shows that the combined weight method is much better in classification in politics category, it means that there are a lot of important words in politics category which are not highly occurred. So, when we use CHI-MCOR, its precise is 3.66% higher than we use CHI method. In fact, when we statistic the top ten of the occurred words in politics category, we find that they are not very high in the total statistics.

5.3 Hierarchical Clustering Experiments

Two feature sets (i.e., word features and concept features) are used in clustering experiments. Fig. 6 is hierarchical clustering for 500 features of character frequency, it shows that the clustering results of word features have no obvious cluster groups and also no wave crests. Fig. 7 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.

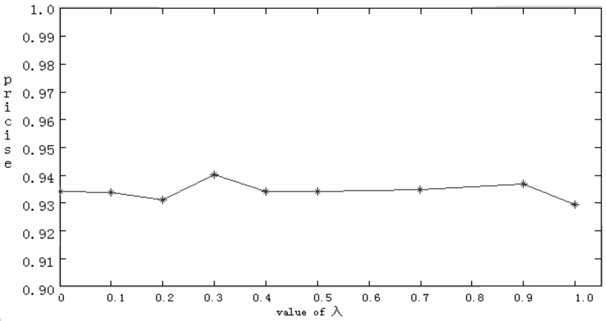

Fig. 4 This is the average precise in CHI-MCOR. The y axis is the average precise, and x axis is the value of λ in the formula which ranges from 0 to 1.

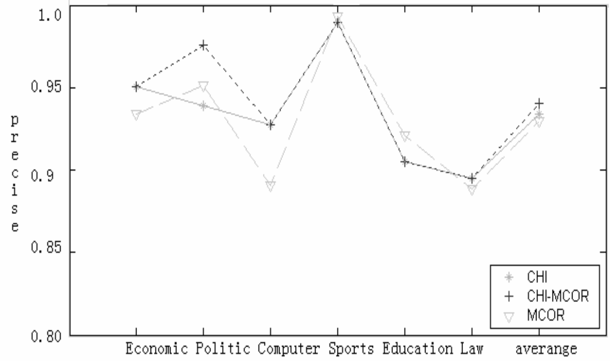

Fig. 5 The precise of the six categories in three weight method, CHI, MCOR and CHI-MCOR when λ is 0.3. The y axis is the classification precise, and x axis is the categories, the last one is the average precise.

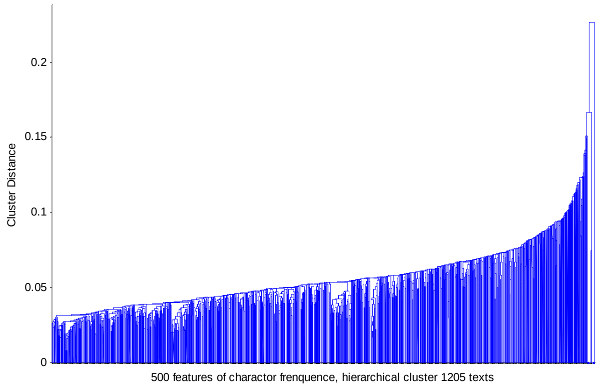

Fig. 6 Hierarchical clusterings for 500 features of character frequency.

5.4 SOM Clustering Experiments

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 which depends on dimension number and distribution of input features, the training time is 4+11 seconds. Fig. 8 and Fig. 9 show the U-matrix (left figure) and D-matrix (right figure) of SOM clustering by using the 1205 texts of 500 word features and 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 cluster borders. There are more cluster borders of high values (black) in Fig. 9, 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 cluster borders. It shows same as hierarchical clustering that the between-cluster distance of concept features is far larger than that of word features.

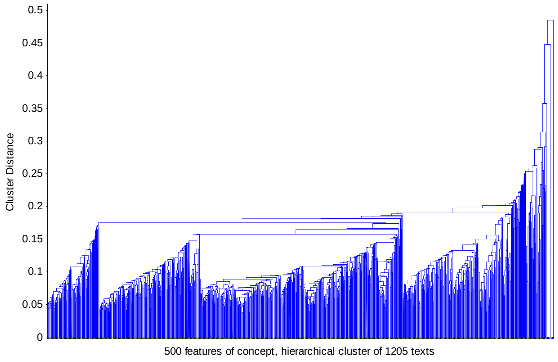

Fig. 7 Hierarchical clusterings for 500 features of concept.

Because the SOM has not distinct cluster boundaries, in order to find and show the borders of the SOM clusters, we use the k-means cluster to find an initial partitioning, the experimental results show that 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. 10 and Fig. 11 are SOM clustering results by using the training the 1205 texts of 500 word features and 500 concept features, the left figures show the Davies-Boulding clustering index [2]; 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 V_{DB} is monotonously decreased with increase of iterative number; the number of clustering groups is also increased. The numbers of the best clusters are 15 and 14 (corresponding to their minimum V_{DB} values) for word features and concept features, respectively.

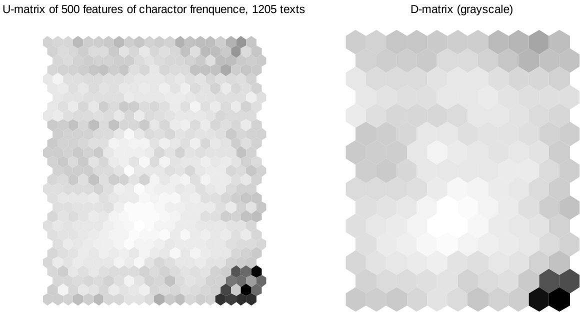

Fig. 8 U-matrix (left figure) and D-matrix (right figure) of SOM, the training data are 1205 texts of 500 word features, SOM's initialization is linear, and batch training algorithm is used in two phases of rough training and fine-tuning.

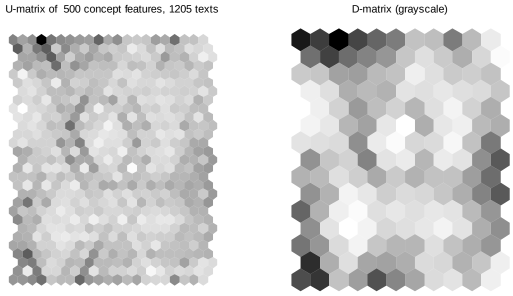

Fig. 9 U-matrix (left figure) and D-matrix (right figure) of SOM, the training data are 1205 texts of 500 concept features, SOM's initialization is linear, and batch training algorithm is used in two phases of rough training and fine-tuning.

The left figures in Fig. 12 and Fig. 13 (which training data are respectively 1205 texts of 500 word features and 500 concept features) are 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. 12 and Fig.13. Small hexagons in Fig. 12 indicate cluster borders (corresponds to large distance between neighboring map units on U-matrix), it shows more small clusters for texts of concept features.

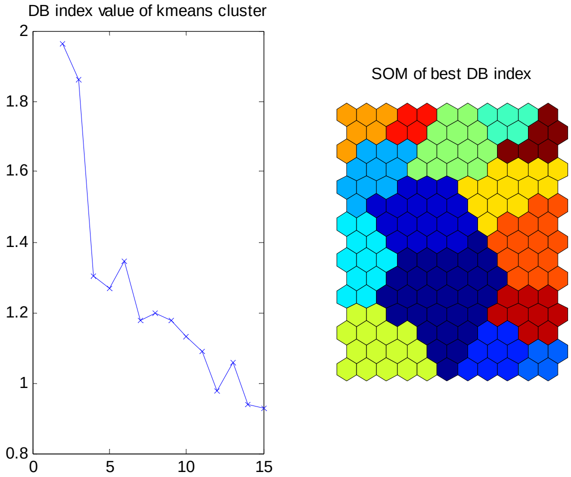

Fig. 10 Davies-Boulding clustering index (left figure), and SOM cluster by color code which is minimized with best clustering (right figure). Training data are 1205 texts of 500 word features.

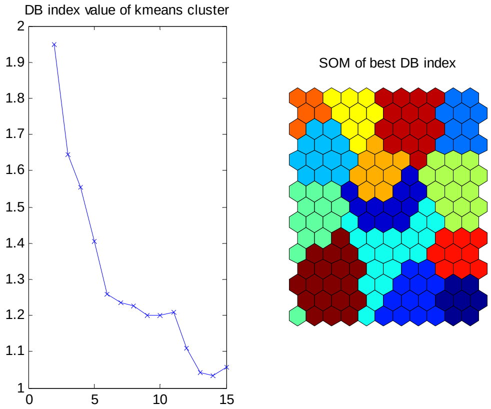

Fig. 11 Davies-Boulding clustering index (left figure), and SOM cluster by color code which is minimized with best clustering (right figure). Training data are 1205 texts of 500 concept features.

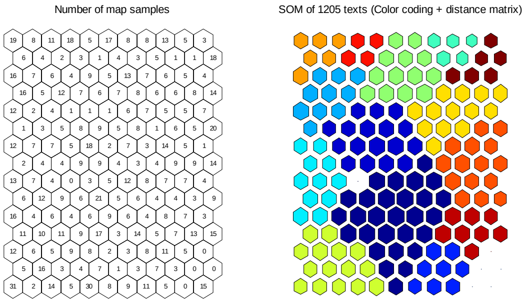

Fig. 12 The distribution of the input data on the map (left figure), the digital in each hexagon is the number of map texts; distance matrix with color codes is shown on the right figure. Training data are 1205 texts of 500 word features.

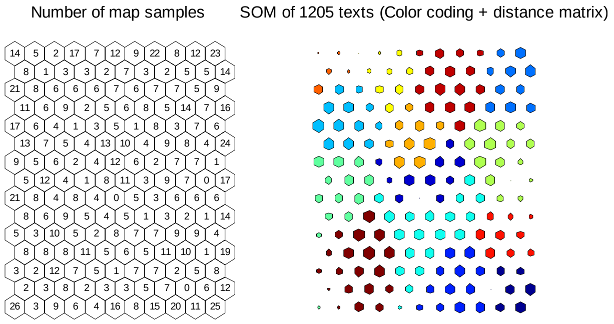

Fig. 13 The distribution of the input data on the map (left figure), the digital in each hexagon is the number of map texts; distance matrix with color codes is shown on the right figure, small hexagons on the D-matrix indicate cluster borders (corresponds to large distance between neighboring map units on U-matrix). Training data are 1205 texts of 500 concept features.

By comparison of hierarchical clustering (Fig. 6 and Fig. 7) and SOM clustering (Fig. 12 and Fig. 13), the results show distinctly easily that between-cluster distance of the texts of concept features is bigger than that of texts of word features, Fig. 6 and Fig. 12 show that word features data nevertheless exhibit some clusters.

By comparing SOM clustering results and artificial classification results, both 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. 12 and Fig. 13, i.e., there are different artificial classes in same output nodes or same color area. The clustering qualities of SOM are evaluated by precision P, recall R and F1. The formulas of precision and recall for class *k* are defined as follows:

$$\Pr ecision_k = \frac{AcorrectNum_k}{AtotalNum_k}.$$

$$\operatorname{Re} call_k = \frac{CorrectNum_k}{TotalNum_k}.$$

The A*correctNum*$_k$ is the number of the documents of the class *k* which are correctly judged by a computer; A*totalNum*$_k$ 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 standard solution. Then the average values of precision, recall and F1can be obtained as the clustering results of SOM:

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

Table 1: SOM Clustering Results

		ECONOMY	POLITY	COMPUTER	SPORT	EDUCATION	LAW	AVERAGE	F1
Word	P	83.0	95.6	93.7	97.2	89.8	87.4	91.12	87.67
	R	82.4	72.3	91.0	92.0	87.6	81.5	84.47	
Concept	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	

Table 1 shows that the clustering performance of concept features is better than that of the word features.

6. Conclusions

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. By setting a shielded level, we can save the word whose DEF is weak in expression and avoid losing important information in concept extraction. When the shielded level is proper, the classification precise is much higher and more stable.

Because there are some dividing words which are not highly occurred but useful in text classification, we use CHI-MCOR method to combine two weight methods together. This method not only selects the highly occurred words, but also selects the dividing word whose occurrence frequency is middle or low. The experimental result shows that CHI-MCOR method is much better than any one of the weight methods. 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, word features data nevertheless exhibit some clusters.

Acknowledgments

This work was supported by Tsinghua University 985 research fund; the excellent young teacher program, Ministry of Education, China; and State Key Lab of Pattern Recognition open fund, Chinese Academy of Sciences.

References

[1]. Q. Zhou, M. Zhao, W. Hu, "The feature selection in Chinese text classification," Journal of Chinese Information Processing, vol.18, no.3, pp.17-21, 2003(in Chinese).

[2]. H. Ji, Z. Luo, M. Wan, et al, "Research on automatic summarization based on concept counting and semantic hierarchy analysis for English texts," Journal of Chinese Information Processing,vol.17, no.12, pp.14-19, 2003.

[3]. C. Li, Z. Luo, Y. Li, "Research on automatic classification of documents based on semantic relativity and concept elativity," Computer Engineering and Application, no.12, pp.106-110, 2003(in Chinese).

[4]. D. Mlademnic, M. Gtobelnik, "Feature selection for unbalanced class distribution and Naïve Bayees," In: Proceedings of the Sixteenth International Conference on Machine learning, pp.258-267, 1999.

[5]. Z. Dong, Q. Dong. The download of Hownet [EB/OL], http://www.keenage.com

[6]. Y. Yang, J. O. Pedersen, "A comparative study on feature selection in text categorization," In: Proceedings of the 14th International Conference on Machine Learning, pp.534-547, 1997

[7]. Y. Li, X. Li, H. Liu, et al., "A novel feature selection method for web pages categorization," Computer Applications, vol.24, no.7, pp.119-203, 2004(in Chinese).

[8]. P. Demartines, J. Herault, "Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets," IEEE Transactions on Neural Networks, vol.8, no.1, pp.148-154, 1997.

[9]. D. Davies, D. Bouldin, "A cluster separation measure," IEEE Transactions on Pattern Analysis and Machine Intelligence - I, vol.2, pp.224-227, 1979.

[10]. T. Kohonen, "The self-organnized map," Proceedings of the IEEE, vol.78, pp.1464-1480, 1990.

[11]. J. Vesanto, J. Alhoniemi, "Clustering of the self-organizing map," IEEE Transactions on Neural Networks, vol.11, no.3, pp.586-600, 2000.

[12]. M. Jiang, H. Cai, B. Zhang, "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, vol.3498, pp.971-976, 2005.

Lin Wang received Bachelor & Master degrees in Dept of Biology, Shandong Teacher University, China, 1986 and 1989, respectively. 1989- 1995 worked as a lecturer in Dept of Biochemistry, Shandong Education University, China; 1995-1998, Researcher in histology & embryology, Medicine College, Shandong University, China; 1998-2000 worked as a lecturer of bio-chemistry, School of Life Science, Tsinghua University, Beijing, China; 2001-2004, Ph.D AG Molekularbiologie, Klinik für Innere Medizin II, Friedrich-Schiller-Universität Jena, Germany. Currently she is an associate professor at School of Electronics & Engineering, Beijing University of Post and Telecommunication, Beijing, her research interests include biomedicine, biotechnology and bioinformatics. About 20 published papers in international journals and conferences recent years.

Minghu Jiang received Bachelor & Master degrees in Dept of Electronics Engineering, Shandong University, China, 1984 and 1989, respectively; 1989-1995 he worked as a lecturer in Dept of Electronics Engineering, Shandong University, China; and received the Ph.D. degree in Signal and Information Processing from Beijing Jiaotong University in 1998. From 1998 to 2000 he worked as postdoctoral fellow at Dept. of Computer, Tsinghua University, Beijing. From 2000 to 2001 he was a postdoctoral researcher in the Dept. of Electrical Eng., K. U. Leuven, Belgium. Currently, he is a professor and director of the lab of computational linguistics, Tsinghua University, Beijing. He is the author or co-author of more than 80 published papers in international journals and conferences. His research interests include neural networks for signal processing, natural language processing.

Shasha Liao received the BS in Computational Linguistics from Tsinghua University in 2003, Beijing. Currently, she is a graduate student of Computational Linguistics, Tsinghua University, Beijing. Her research interest is natural language processing.

Yinghua Lu is a full professor in School of Electronics & Engineering, Beijing University of Post and Telecommunication, Beijing, China. He is the author or co-author of more than 160 published papers in international journals and conferences. His research interests include signal processing, biomedicine engineering and biotechnology.