



## A New Method of Hierarchical Text Clustering Based on Lsa-Hgsom

Jianfeng Wang, Lina Ma, Xinye Li, Yangxiu Zhou & Dong Qiao

Technology College, North China Electric Power University

No 282, Rei Xiang Street, Baoding, China

Tel: 86-312-752-3461 E-mail: wjf611@yahoo.com.cn

*This research was supported by Young Teachers Research Foundation of North China Electric Power University No.93222805*

### Abstract

Text clustering has been recognized as an important component in data mining. Self-Organizing Map (SOM) based models have been found to have certain advantages for clustering sizeable text data. However, current existing approaches lack in providing an adaptive hierarchical structure within in a single model. This paper presents a new method of hierarchical text clustering based on combination of latent semantic analysis (LSA) and hierarchical GSOM, which is called LSA-HGSOM method. The text clustering result using traditional methods can not show hierarchical structure. However, the hierarchical structure is very important in text clustering. The LSA-HGSOM method can automatically achieve hierarchical text clustering, and establishes vector space model (VSM) of term weight by using the theory of LSA, then semantic relation is included in the vector space model. Both theory analysis and experimental results confirm that LSA-HGSOM method decreases the number of vector, and enhances the efficiency and precision of text clustering.

**Keywords:** Text Clustering, Hierarchical GSOM, Latent Semantic Analysis, Vector Space Model

### 1. Introduction

With the popularization and application of Internet network has become an important part of the people's working and living, and various search engines have been an indispensable tool to retrieve the necessary resources for the people. However, the Internet search engine can often find thousands of search results. Even if some useful information is obtained, it is often mixed with a lot of "noises" to waste the users' time and money. Therefore, in order to efficiently and economically retrieve the resource subset relevant to the given search request and with the appropriate number, the Text clustering is performed and becomes one of important and hot research fields in data mining (R.D.Lawrence,1999,pp.171-195).

Text clustering is different from Text classification. The latter has them for each category while Text clustering has no category annotates in advance. The Text clustering is to divide the Text sets into several clusters according to the Text contents, and requires the similarity of the Text contents in the clusters as great as possible and that of different clusters as small as possible. It can organize the Web Text effectively, but also form a classification template to guide the classification of the Web Text. Therefore, the Text clustering is an important content in the domain of data mining, and also acts as a very important role in text mining. The general procedure of the text clustering methods is as follows. Firstly, the documents to be clustered are transformed into some sets of terms, and term weights are assigned to each term of the sets, then some term weights constitute a feature vector that represents a text. In fact, text clustering means text contents clustering. However, the term sets can not concern with the text contents in clustering process. Therefore, a way of improving text clustering effect is that clustering of documents is based on text conception (or semantic content). The existing methods of text clustering can obtain a single clustering structure; however, the result can not show hierarchical relation among categories. In fact, people need to understand the hierarchical relation among categories. In order to overcome this defect, we present a new method of hierarchical text clustering called LSA-HGSOM method. The method applies the theory of LSA to construct a VSM (Vector Space Model), and achieves text clustering through conception statistic computation. Therefore, the method advances the speed and precision of text clustering. Moreover, the clustering method can achieve automatically hierarchical text clustering through a hierarchical GSOM method (called HGSOM).

## 2. The theory of LSA

### 2.1 Term Matrix

LSA (Latent Semantic Analysis) (S.T.Dumais,1988,pp.281-285) is one of the most popular linear document indexing methods which produce low dimensional representations using word co-occurrence which could be regarded as semantic relationship between terms. LSA aims to find the best subspace approximation to the original document space in the sense of minimizing the global reconstruction error (the Euclidean distance between the original matrix and its approximation matrix). It is fundamentally based on SVD (Singular Value Decomposition) and projects the document vectors into the subspace so that cosine similarity can accurately represent semantic similarity. Given a term-document matrix  $X = [x_1, x_2, \dots, x_n] \in R^m$  and suppose the rank of  $X$  is  $r$ , LSA decomposes  $X$  using SVD as follows:

$$X = U \Sigma V^T \quad (1)$$

where  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$  and  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$  are the singular values of  $X$ .  $U = [u_1, \dots, u_r]$  and  $u_i$  is called the left singular vector.  $V = [v_1, \dots, v_r]$  and  $v_i$  is called the right singular vector. LSA uses the first  $k$  vectors in  $U$  as the transformation matrix to embed the original documents into a  $k$ -dimensional space.

### 2.2 Singular Value Decomposition

After the matrix  $X$  is established, we can acquire an approximate matrix  $X_k$  of the matrix  $X$  with  $k$  orders, where  $k < \min(m, n)$ . By the singular value decomposition (G. W. Furnas, 1988, pp.36-40), a matrix  $X$  can be denoted as a product of three matrices.

$$X = U \Sigma V^T \quad (2)$$

In formula (2),  $U$  and  $\Sigma$  denote the left and the right singular vector matrices of the matrix  $X$  respectively; the diagonal matrix  $\Sigma$  consists of singular values of the matrix  $X$  according to the arrangement with descending order. We select the foremost  $k$  maximal singular values of the matrix  $X$ , and establish an approximate matrix  $X_k$  with  $k$  order.

$$X_k = U_k \Sigma_k V_k^T \quad (3)$$

In formula (3),  $U_k$  and  $V_k$  are orthogonal vectors.  $X_k$  denotes approximately the term vector matrix  $X$ , the row vectors of  $U_k$  represent the term vectors, and the row vectors of  $V_k$  represent the document vectors. After using singular value decomposition and selecting approximate matrices of  $k$  orders, the model of LSA acquires some good effects as follows. For one thing, the disadvantageous factors in original term matrix are decreased. Moreover, the semantic relation between terms and documents becomes more obvious. In addition, the dimension of VSM is decreased greatly, and so the speed of clustering is advanced. In short, through the process of LSA, the VSM of documents has the following merits.

The dimension of VSM is decreased greatly, and so the speed of clustering is advanced.

## 3. The theory of GSOM

### 3.1 The Self-Organizing Map (SOM)

SOM is an unsupervised neural network model that maps high-dimensional input space to low-dimensional output space. When the resulting map is a two-dimensional topology, the intuitive visualization provides good exploration possibilities. The drawback of this approach is the pre-fixed structure of the output space and lack of providing hierarchical relations between the input spaces. (A. Rauber, 1999, pp. 302-311)

### 3.2 The Growing Self Organizing Map (GSOM)

The GSOM algorithm is composed of three phases, initialization, growing and smoothing. Soon after the smoothing phase, the generated map can be queried and the input data vectors clustered (A. Hsu, 2003, pp. 2131-2140).

#### 3.2.1 Initialization phase:

Initialize the weight vectors of the starting nodes (usually four) with random numbers between 0 and 1.

Calculate the growth threshold (GT) for the given data set of dimension  $D$  according to the spread factor (SF) using the formula  $GT = -D \times \ln(SF)$

#### 3.2.2 Growing Phase:

a) Present input to the network.

b) Determine the weight vector that is closest to the input vector mapped to the current feature map (winner), using Euclidean distance. This step can be summarized as: find  $q'$  such that  $|v - w_{q'}| \leq |v - w_q| \forall q \in N$  where  $v$ ,  $w$  are the

inputs and weight vectors respectively,  $q$  is the position vector for nodes and  $N$  is the set of natural numbers.

c) The weight vector adaptation is applied only to the neighborhood of the winner and the winner itself. The neighborhood is a set of neurons around the winner, but in the GSOM the starting neighborhood selected for weight adaptation is smaller compared to the SOM (localized weight adaptation). The amount of adaptation (learning rate) is also reduced exponentially over the iterations. Even within the neighborhood, weights that are closer to the winner are adapted more than those further away. The weight adaptation can be described by

$$w_j(k+1) = \begin{cases} w_j(k) & \text{if } j \notin N_{k+1} \\ w_j(k) + LR(k) \times (x_k - w_j(k)) & \text{if } j \in N_{k+1} \end{cases} \quad (4)$$

Where the Learning Rate  $LR(k)$ ,  $k \in N$  is a sequence of positive parameters converging to zero as  $k \rightarrow \infty$ .  $w_j(k)$ ,  $w_j(k+1)$  are the weight vectors of the node  $j$  before and after the adaptation and  $N_{k+1}$  is the neighborhood of the winning neuron at the  $(k+1)$ th iteration. The decreasing value of  $LR(k)$  in the GSOM depends on the number of nodes existing in the map at time  $k$ .

- d) Increase the error value of the winner (error value is the difference between the input vector and the weight vectors).
- e) When  $TE_i \geq GT$  (where  $TE_i$  is the total error of node  $i$  and  $GT$  is the growth threshold). Grow nodes if  $i$  is a boundary node. Distribute weights to neighbors if  $i$  is a non-boundary node.
- f) Initialize the new node weight vectors to match the neighboring node weights.
- g) Initialize the learning rate (LR) to its starting value.
- h) Repeat steps 2 – 7 until all inputs have been presented and node growth is reduced to a minimum level.

### 3.2.3 Smoothing phase.

- a) Reduce learning rate and fix a small starting neighborhood.
- b) Find winner and adapt the weights of the winner and neighbors in the same way as in growing phase.

The growth threshold is based on the number of dimensions of the dataset and the spread factor (SF). SF is a predetermined value in the range 0-1, with zero allowing least spread and one, maximum spread. A limited spread with a smaller SF value should ideally be the starting map. Once significant clusters are identified, they can be used as the basis for further analysis with a higher SF value.

## 4. Text clustering method based on LSA--HGSOM

### 4.1 Conception of Text Clustering

Text clustering is the process of assigning the documents in a document base to different categories, and is a typical machine learning problem with no supervising. A species is some groups of documents. Documents within one species are more similar than those among different species. Therefore, the aim of text clustering is to group the documents: minimizing the similarity among different species and maximizing the similarity within a species.

Clustering analysis is the process that assigns similar documents to the same categories through computing the degree of similarity among all documents. By the singular value decomposition, the row vectors of  $V$  are the vectors of texts. Therefore, we apply the row vectors of  $V$  to calculating the degree of similarity among documents. The degree of similarity is generally denoted by cosine distance, which is defined as follows:

$$Sim(i, j) = \frac{\sum_{m=1}^k W_{im} \times W_{jm}}{\sqrt{\sum_{m=1}^k (W_{im})^2 \times \sum_{m=1}^k (W_{jm})^2}} \quad (5)$$

Where  $Sim(i, j)$  is the degree of similarity between text  $i$  and  $j$ ; where  $W_{im}$ , and  $W_{jm}$  denote the values of the rows  $i$  and  $j$  of the column  $m$  in the matrix  $V$  respectively.

### 4.2 Dimensionality Reduction

Reduction of the data dimensionality may lead to significant savings of computer resources and processing time. However the selection of fewer dimensions may cause a significant loss of the document local neighborhood information. Due to this compromise, we have chosen to use the popular and well studied singular value decomposition. SVD is used to rewrite an arbitrary rectangular matrix, such as a Markov matrix, as a product of three other matrices:  $X = U \Sigma V^T$ . As a Markov matrix is symmetric, both left and right singular vectors ( $U$  and  $V$ ) provide a mapping from the document space to a newly generated abstract vector space. The elements  $(\lambda_0, \lambda_1, \dots, \lambda_{r-1})$  of the diagonal matrix  $S$ , the singular values, appear in a magnitude decreasing order. One of the more important theorems of

SVD states that a matrix formed from the first  $n$  singular triplets  $\{U_i, \lambda_i, V_i\}$  of the SVD (left vector, singular value, right vector combination) is the best approximation to the original matrix that uses  $n$  degrees of freedom. The technique of approximating a data set with another one having fewer degrees of freedom, known as dimensional reduction, works well, because the leading singular triplets capture the strongest, most meaningful, regularities of the data. The latter triplets represent less important, possibly spurious, patterns. Ignoring them actually improves analysis, though there is the danger that by keeping too few degrees of freedom, or dimensions of the abstract vector space, some of the important patterns will be lost.

After reducing the dimension, documents are represented as  $n$ -dimensional vectors in the diffusion space, and can be clustered by using HGSOM.

#### 4.3 HGSOM Clustering

Hierarchical clustering techniques are categorized into agglomerative (bottom-up) and divisive (top-down) approaches (A.Hsu,2003,pp.2131-2140). Agglomerative clustering starts with one point clusters and recursively merges two or more similar clusters until all the clusters are encapsulated into one final cluster. Divisive clustering considers the entire dataset as one cluster and then recursively splits the most appropriate cluster until a stopping criterion is achieved. For details on clustering algorithms refer to. The hierarchical clustering model presented in this section builds a hierarchy of clusters in a novel manner. It does not follow a traditional bottom-up or top-down approach, but using GSOM as the basis and utilizing its spread factor. Since the spread factor takes the value between 0 and 1, to avoid missing any significant sub groupings, a set of values across the whole range (0-1) are initialized.

The GSOM uses a threshold value,  $GT$ , to decide when to initiate new node growth (Amarasiri,2000,pp.601-614).  $GT$  will decide the amount of spread of the feature map to be generated. Therefore, if only an abstract picture of the data is required, a large  $GT$  will result in a map with a fewer number of nodes (A.Hsu,2003,pp.2131-2140). Similarly, a smaller  $GT$  will result in the map spreading out more. Node growth in the GSOM is initiated when the error value of a node exceeds the  $GT$ . The total error value for node  $i$  is calculated as:

$$TE_i = \sum_{H_i} \sum_{j=1}^D (x_{i,j} - w_j)^2 \tag{6}$$

where  $H_i$  is the number of hits to the node  $i$  and  $D$  is the dimension of the data.  $X_{i,j}$  and  $w_j$  are the input and weight vectors of the node  $i$ , respectively. For a boundary node to grow a new node, it is required that

$$TE_i \geq GT \tag{7}$$

The  $GT$  value has to be experimentally decided depending on the requirement for the map growth. As can be seen from (7), the dimension of the data set will make a significant impact on the accumulated error (TE) value, and as such will have to be considered when deciding the  $GT$  for a given application. Since  $X_{i,j} \geq 0$ ,  $W_j \leq 1$ , the maximum contribution to the error value by one attribute (dimension) of an input would be,

$$\max |x_{i,j} - w_j| = 1 \tag{8}$$

Therefore, from (7)

$$TE_{\max} = D \times H_{\max} \tag{9}$$

where  $TE_{\max}$  is the maximum error value and is the maximum possible number of hits. If  $H_i$  is considered to be the number of hits at time (iteration)  $t$ , the  $GT$  will have to be set such that

$$0 \leq GT < D \times H(t) \tag{10}$$

Therefore,  $GT$  has to be defined based on the requirement of the map spread. It can be seen from (10) that the  $GT$  value will depend on the dimensionality of the data set as well as the number of hits. Thus, it becomes necessary to identify a different  $GT$  value for data sets with different dimensionality. This becomes a difficult task, especially in applications such as data mining, since it is necessary to analyze data with different dimensionality as well as the same data under different attribute sets. It also becomes difficult to compare maps of several datasets since the  $GT$  cannot be compared over different datasets. Therefore, the user definable parameter is introduced. The  $SF$  can be used to control and calculate the  $GT$  for GSOM, without the data analyst having to worry about the different dimensions. The growth threshold is defined as

$$GT = D \times f(SF) \tag{11}$$

where  $SF \in R, 0 \leq SF \leq 1$ , and  $f(SF)$  is a function of  $SF$ , which is identified as follows.

The total error  $TE_i$  of a node  $i$  will take the values

$$0 \leq TE_i \leq TE_{\max} \tag{12}$$

where  $TE_{\max}$  is the maximum error value that can be accumulated. This can be written as

$$0 \leq \sum_H \sum_{j=1}^D (x_{i,j} - w_j)^2 \leq \sum_{H_{\max}} \sum_{j=1}^D (x_{i,j} - w_j)^2 \tag{13}$$

Since the purpose of the GT is to let the map grow new nodes by providing a threshold for the error value, and the minimum error value is zero, it can be argued that for growth of new nodes,

$$0 \leq GT \leq \sum_{H_{\max}} \sum_{j=1}^D (x_{i,j} - w_j)^2 \tag{14}$$

Since the maximum number of hits ( $H_{\max}$ ) can theoretically be infinite, (14) becomes  $0 \leq GT \leq \infty$ . According to the definition of spread factor, it is necessary to identify a function  $f(SF)$  such that  $0 \leq D \times f(SF) \leq \infty$ .

A function that takes the values  $0 \rightarrow \infty$ , when  $x$  takes the values  $0$  to one, is to be identified. A Napier logarithmic function of the type  $y = -a \times \ln(1 - x)$  is one such equation that satisfies these requirements. If  $\mu = 1 - SF$  and

$$GT = -D \times \ln(1 - \eta) \tag{15}$$

then

$$GT = -D \times \ln(SF) \tag{16}$$

Therefore, instead of having to provide a GT, which would take different values for different data sets, the data analyst can now provide a value - SF, which will be used by the system to calculate the GT value depending on the dimensions of the data. This will allow HGSOM to be identified with their spread factors and can form a basis for comparison of different maps.

During cluster analysis, it may be necessary (and useful) for the analyst to study the effect of removing some of the attributes (dimensions) on the existing cluster structure. This might be useful in confirming opinions on non-contributing attributes on the clusters. The spread factor facilitates such further analysis since it is independent of the dimensionality of the data. This is very important, as the growth threshold depends on the dimensionality.

**5. Experiments**

We applied the theory of LSA method to construct the VSM, and applied the HGSOM network to achieve text clustering. We collected 500 documents to carry on the text clustering. These documents were classified into three large-scale species and six child species: Education (Includes campus (CAM) and high school education (HSE)), Economics (Includes industrial (IND) and agriculture (AGR) economics), and Medicine (Includes clinic (CLI) and nurse (NUR)). After the feature selection, we obtained 1863 feature words.

*5.1 Experiment 1*

We applied directly the words matrix D to carry on the text clustering, and the number of input node was 1863. In this case we could not acquire satisfactory result using HGSOM. The training speed of the network is very slow, and the result of text clustering is not correct. We think that the wrong result comes of too many input nodes. So we could not acquire good result using HGSOM network.

*5.2 Experiment 2*

In view of the result of the experiment 1, we choose k=785 through the process of LSA. After that, the number of input nodes is reduced to 785. We applied the HGSOM clustering algorithm to the documents clustering. The speed of training is improved greatly, and the results of clustering are shown in the table 1 and table 2. We applied average accuracy (AA%) (Jiang Ning,2002,pp.5) to evaluate the results of text clustering.

In view of the result of the experiment 1, we choose k=785 through the process of LSA. After that, the number of input nodes is reduced to 785. We applied the HGSOM clustering algorithm to the documents clustering. The speed of training is improved greatly, and the results of clustering are shown in the table 1 and table 2. We applied average accuracy (AA%) (Jiang Ning,2002,pp.5) to evaluate the results of text clustering.

In table 1 and table 2, the clustering results are satisfactory; therefore, HGSOM method is feasible for hierarchical text clustering. From table 1 and table 2, we can see that the results of table 1 are better than the results of table 2. The reason of acquiring this result is that the documents have more similar feature words among sub-layers than those documents among large-scale species, which impacts on the result of text clustering.

**6. Conclusions**

We can draw the conclusions from above experiments as follows:

- (1) The present LSA-HGSOM method is feasible for text clustering.

- (2) LSA-HGSOM method can achieve automatically hierarchical text clustering, and overcome the shortcomings of traditional methods.
- (3) LSA-HGSOM network is limited for text clustering. If there are many input nodes, we can not acquire good result.
- (4) LSA-HGSOM method can enhances the efficiency and precision of text clustering.

In this paper, we proposed a new text clustering method LSA-HGSOM. In this method, it firstly makes preprocess of texts, and introduced LSA theory to improve the precision of clustering and reduce the dimension of feature vector. Then it used HGSOM to execute the clustering to the texts, In the experiment, the result shows that LSA-HGSOM method would get a better effect in the text clustering.

### References

- A.Hsu, S.L.Tang & S.K.Halgamuge. (2003). An unsupervised hierarchical dynamic self-organising approach to cancer class discovery and marker gene identification in microarray data, *Bioinformatics*, vol.19, pp.2131-2140.
- A.Rauber & D.Merkl. (1999). Using self-organizing maps to organize document archives and to characterize subject matter: How to make a map tell the news of the world, *In DEXA'99: Proceedings of the 10th International Conference on Database and Expert Systems Applications*. London, UK: Springer-Verlag, pp.302-311.
- Amarasiri, R., & Alahakoon, D. (2000). *Applying Dynamic Self Organizing Maps for Identifying Changes in Data Sequences*. *IEEE Transactions on Neural Networks, Special Issue on Knowledge Discovery and Data Mining*, 11(3).pp.601-614.
- G.W.Furnas. (1988). Information retrieval using a singular-value-decomposition model of latent semantic structure. [c] *In Proceedings of SIGIR'88*, 1988, pp.36-40.
- Jiang Ning & Shi Zhong-zhi. (2002). Bayesian posteriori model selection for text clustering, *Journal of computer research and development*, pp.5.
- R. D. Lawrence, G.S.Almasi & H.E.Rushmeier. (1999). A scalable parallel algorithm for selforganizing maps with applications to sparsedata mining problems, *Data Mining and Knowledge Discovery*, 3,171-195.
- S.T.Dumais. (1988). Using latent semantic analysis to improve information retrieval. *In CHI'88 Proceedings*. 281-285.

Table 1. The clustering result of the first layer by HGSOM method

	Medicine	Education	Economics
(AA)%	88.5	90.2	86.3

Table 2. The clustering result of the sub-layer by HGSOM method

	CLI	NUR	CAM	HSE	IND	AGR
(AA)%	79.6	79.1	84.1	82.9	85.4	84.9