USING SOM NEURAL NETWORK IN TEXT INFORMATION RETRIEVAL

J. Mehrad, Ph.D.
President of RLST
email: dean@srlst.com

S. Koleini, M.S.
Head, Dept. of Information Technology, RLST
email: koleini@srlst.com

Abstract - With the increase of the volume of information and the progress in technology, the deficiency of traditional algorithms for fast information retrieval becomes more clear. When large volumes of data are to be handled, the use of neural network as an Artificial Intelligent technique is a suitable method to increase the information retrieval speed. Neural networks present a suitable representation of knowledge in retrieval applications. The nodes in neural network, present the items of information retrieval such as keywords, authors,... and links are used for data transfer between layers. Finally, it results in a network output that is a retrieval document. In this article, the use of SOM neural network for data clustering is shown. A model of SOM neural network for a sample information retrieval from INIS database is implemented.

Keywords - Information, Text Retrieval, Neural Network, Self Organizing Map, Clustering Algorithm.

INTRODUCTION

Retrieval strategies describe the similarity between a Query and a Document. These strategies are based on the relevancy between Queries and Documents. A retrieval strategy is an algorithm that uses Query (Q) and Document collection (D1, D2,...., Dn). To calculate the SC(Q,Di) (Similarity Coefficient) for all documents, one of the effective strategies in information retrieval is to use neural network systems. This strategy includes a collection of network neurons that are activated in a query and during the document retrieval period [4]. A neural network contains nodes and links, which are defined according to the system input and output values. In various models of neural network, information is presented as weighted network. In spite of traditional techniques of information retrieval, neural network models operate as self processors without using any other external programs [1].

The network, with its talented behavior, in local interaction that occurs simultaneously in all parts of network performs the information processing as well.

In traditional information retrieval model, external processing on data structures, usually access to all network rules and processing is considered to be a sequential task. It seems that neural network calculation in comparison to vector space model and probability model, better fits the traditional information retrieval models [5]. Using neural

network in information retrieval has some advantages such as:

1. When the requested data (keywords) do not exist in the document collection, we can use the neural network to retrieve the information proximity around the required information.
2. The information can be classified with common patterns.

The SOM algorithm was introduced in1981 and attracted a great deal of interest among researchers and practitioners in a wide variety of fields. The SOM has been analyzed extensively, a number of variants have been developed and perhaps it has been applied extensively within fields ranging from Engineering Science to Medicine, Biology and Economics. There were 5384 articles on SOM from 1981 till 2002 and it shows the value of this subject [11]. Laaksonen and Koskela [9] show that if the data contain semantically related object groupings or classes, subsets of vectors belonging to such user-defined classes can be mapped on the SOM by finding the best matching unit for each vector in the set. The distribution of the data vectors over the map forms a two-dimensional discrete probability density. Even from the same data, qualitatively different distributions can be obtained by using different feature extraction techniques. They used such feature distributions to compare different classes and different feature representations of the data in the context of our content-based image retrieval system PicSOM. With the WEBSOM method, a textual document collection may be organized onto a graphical map display that provides an overview of the collection and facilitates interactive browsing. Interesting documents can be located on the map using a content-directed search. Each document is encoded as a histogram of word categories which are formed by the self-organizing map algorithm based on the similarities in the contexts of the words. The encoded documents are organized on another self-organizing map, a document map, on which nearby locations contain similar documents. Special consideration is given to the computation of very large document maps, which is possible with general-purpose computers if the dimensionality of the word category histograms is first reduced with a random mapping method and if computationally efficient algorithms are used in computing the SOM [7,10]. Ultsch, in his work, shows that the application of SOM for data mining that is called Neuronal Data Mine uses emergent feature maps and knowledge conversion [12].

RESEARCH METHOD

The text data from INIS (International Nuclear Information System) CD-Rom were used as our database. The objective of the model was to retrieve information with certain subjects. Our subject is "cross-section" in three parts as follows:

1. Elastic scattering cross-section
2. Absorption
3. Fission

Fifty documents for each subject are selected. As a result, a document collection with 150 documents is obtained. The objective is the classification of documents into three types (i.e. Elastic, Absorbtion and Fission) with neural network clustering algorithm. Knowing that in *'Fission'* subject, the term of *"neutron"*, in *"Elastic"* subject, the term of *"proton"* and finally in "absorbtion" subject the term *"electron"* and *"neutron"* have occurred more than the other keywords, the weights of documents are calculated and a program is written to preprocess and code the text document into numeric vectors. To determine the weight vector for each document, the keyword frequency in document is used. In other words, document weight is defined as keyword frequency (t_k) in document (d_j). To calculate the weights, the following equation is used [4]:

$$W_{ik} = \frac{(tf_{ik} . \log(N/n_k))}{\sqrt{\sum_{j=1}^{t} (tf_{ij})^2 . \log(N/n_j))^2}} \qquad (1)$$

In this equation: tf_{ik} is keyword frequency in document d_i, N is document number and N_k refers to the documents that include the word or keyword (t_k). The above parameters for 150 sample documents will be determined.

This algorithm is performed by first associating a node in the network for each cluster. Each node then computes (in parallel) a measure of similarity between the existing document and the centroid that represents the cluster associated with the node. First, a similarity coefficient is computed between the incoming document X and the existing cluster centroids. The input nodes of the neural network correspond to each cluster. If the similarity coefficient, s_1, is higher than a threshold, S_{1ovg}, the input node is activated. It then loops back to itself after a small recalculation to participate in a competition to add X to the cluster. Nodes that are not close enough to the incoming document are deactivated.

A new pass then occurs for all the nodes that won the first round and the similarity coefficient is computed again. The process continues until only one cluster passes the threshold. At this point, a different similarity coefficient is computed, s_2, to ensure the winning cluster is reasonably close to the incoming document. If it is close enough, it is added to the cluster and the centroid for the cluster is updated. Otherwise, a new cluster is formed with the incoming document.

The self organized map (SOM) learning process is considered as competition learning (see Figure1).

step0 initialize weights w_{ij}
set topological neighborhood parameters
set learning rate parameter
step 1 while stopping condition is false do step 2-8
step2 for each input vector x , do step 3-5
step3 for each j, compute:
$d(j) = \sum_i (w_{ij} - x_i)^2$
step4 find index j such that d(j) is a minimum
step 5 for all units j whitin a specified neighborhood of j and for all i:
w_{ij} (new)= w_{ij} (old)+ $\alpha[x_i - w_{ij}$ (old)]
step6 update learning rate
step7 reduce radius of topological neighborhood at specified time
step8 test stopping condition

Figure 1: The SOM learning algorithm.

The main idea of competition learning is to adjust the test cluster (c) of the network with high level activity according to the selected random inputs. Output activation level is according to Euclidean, the distance between clusters weight vector (mc) and the input. Vector space model is used for representing of data with vectors that determin the weight *w*.
In this case,
$\sum (w_i)^2=1$. The similarity between two terms is determined by vector cosine similarity as follows [1]:

$$\frac{(w,v)}{\|w\|^2\|v\|^2} = \frac{(w.v)}{\|w\|\|v\|} = \frac{(w.v)}{(1)(1)} = \sum W_i V_i \qquad (2)$$

In this measurement, the angle between two vectors is determined. A vector for each document according to four type keywords is constructed. The algorithm for this processing is shown in Figure 2.

The following four steps are considered for learning process:

1. Select the random input x (t) .
2. Calculate the distance between weight vector and input vector using this equation:

$$D_i(t) = \|x(t) - m_i(t)\| \qquad (3)$$

In this equation, m_i refers to the weight vector *i* and $\|.\|$ is Euclidean vector norm.

For N training epochs
For each training documents:
    Word_index=0
Neighbor_ Word_index= Word_index+1
Adjust context vector for word(Word_index) and word(neighbor_Word_index)
D = w1-w2 where
w1= vector for word(Word_index)
w2 = vector for word (neighbor_Word_index)
$w_1 (k+1)= w_1(k)- \mu_w k_1 d$ *where*
μ_w *= learning rate for word neighbor adjustments*
$k_1=(w1\_update*((neighbor\_Word\_index) - (Word\_index)))^{-1}$
$w_2 (k+1)= w_2 (k)+ \mu_w k_2 d$ *where*
$k_2 =(w2\_update*((neighbor\_Word\_index) - (Word\_index)))^{-1}$
normalize w1 and w2
if ((neighbor_Word_index) – (Word_index)) < max_neighbor_word
increment neighbor_Word_index
else if not done with all words in document
    increment Word_index
calculate vector for document
d= w-v where
    w= vector for the word
    v=vector for the entire document
$w(k+1)=w(k)- \mu_d d$
    where μ_d *is learning rate for word-to-document adjustment* ($\mu_d << \mu_w$)
renormalize *w*

Figure 2: The project's algorithm.

3. The winner cluster is obtained by $C : \| x(t) - m_i(t) \| = \min_i (D_i(t))$
4. Adjust the weight vectors in the neighborhood of winner cluster.

The learning rule for SOM neural network is used as follows:

$$m_i(t+1) = m_i(t) + \mu(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)]$$

In this equation, $\mu(t)$ is a coefficient reduced by time and $h_{ci}(t)$ is the neighboring function that is symmetric in the winner neighboring radius.

In this research, simple Gaussian neighboring function is used as follows:

$$h_{ci}(t) = \left(\frac{\|r_c - r_i\|}{2.\delta(t)^2} \right) \qquad (6)$$

The value of δ is reduced by increasing the learning time t. The variable of r is the neighboring radius. With decreasing learning rate and the neighboring radius, learning process goes to its steady state. When the changes in weight vectors are reduced, steady state is obtained. However, it is possible that learning process may stop whenever each input data is assigned to the same cluster repeatedly.

RESULTS

This research is implemented based on the infrastructure provided by SOM Toolbox[1] for MATLAB software (Matlab version 7 is used in this research). By using equation 1, weight document vectors are determined. Since SOM algorithm is based on Euclidean distance, the range of variables are important for dedicating a variable to a specific cluster. As usual, the variables are normalized and then the network is trained with normalized data.

50 sample data based on 3 classes related to the *cross-section* topic were analyzed. Their map is illustrated in Figure 2. As could be seen from Figure 2, the range of normal variable values for each keyword is recognized. In Figure 3, U matrix[2] indicates the neighborhood distance and determines the cluster structure in SOM network.

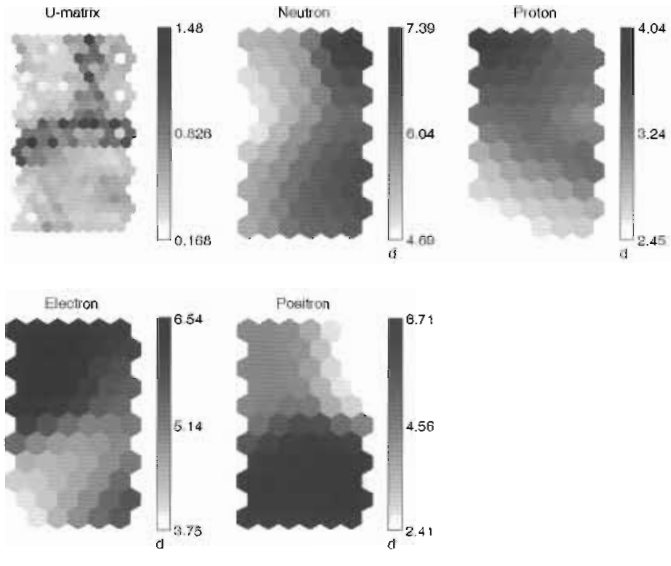

Figure 3: Variable values in each map with U matrix.

To calculate the U matrix, all or parts of the network variables are used. In this case, all network variables are used since the number of variables were few.

In this matrix, high values indicate the neighboring distance between clusters and therefore it determines the cluster limits. Using the color column diagram the value of each color is recognized and the first value indicates the variable value in the SOM network. The U matrix and the labels are illustrated in Figure 4. The cluster classification is shown in this Figure.

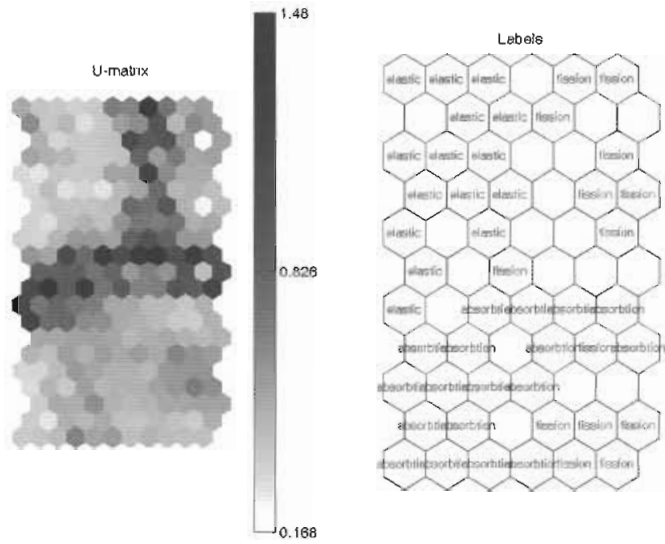

Figure 4: The U matrix with clusters.

The distance matrix based on the surface representation is shown in Figure 5. The mean distance for neighboring neuron in the map is indicated by the value of Z axis. This matrix is very close to U matrix. Matrix topology relation with surface presentation is easily recognized in this Figure.

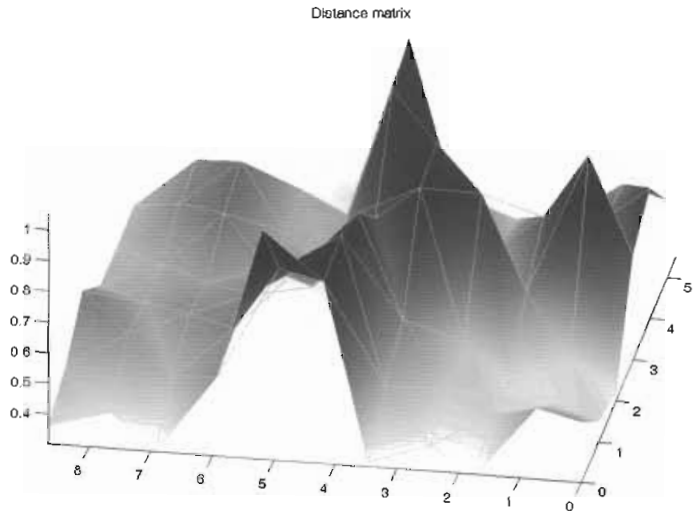

Figure 5: Surface representation of distance matrix.

The scheme of SOM network with its neuron at 3 dimensional spaces is illustrated in Figure 6.

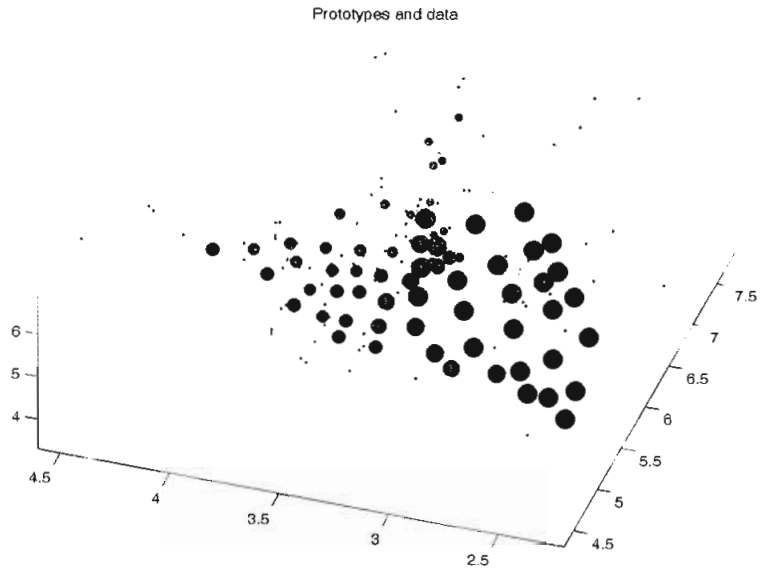

Figure 6: SOM network.

The SOM network with 3 dimensional presentation of data is shown in Figure 7.

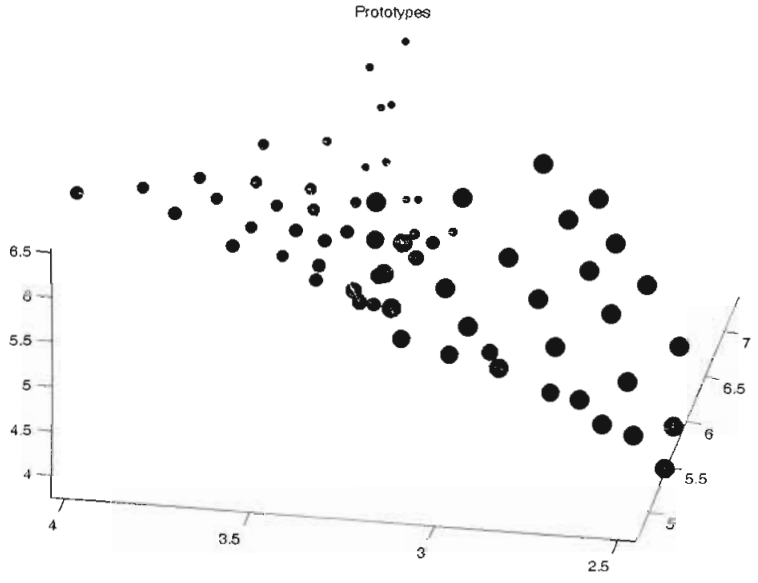

Figure 7: SOM network with real data.

For system evaluation, precision and recall are calculated according to their definitions. Precision is the ratio of the number of relevant documents retrieved and recall is the ratio of the number of relevant documents retrieved to the total number of documents in the collection that are believed to be relevant. In this system, the value of precision is 0.25 and the value of recall is 0.92 as shown in Figure 8. Since the recall value is close to 1, it means that the performance of this system is acceptable.

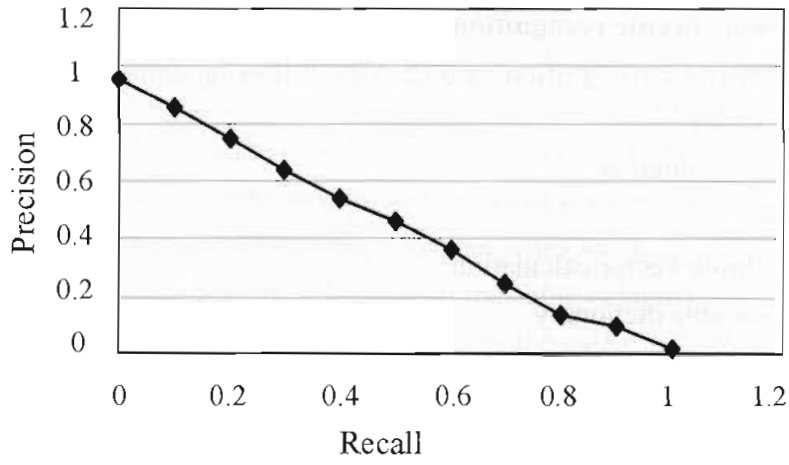

Figure 8: Recall and precision curve.

CONCLUSIONS

Information visualization method is concerned with graphically representing complex, abstract data domains to facilitate knowledge extraction from very large data sets. It is recognized as having the potential to enable better understanding of a complex system, and to allow the discovery of information that might be remained unknown [8].

The SOM is used as a combination of clustering and projection techniques for feature extraction, visualization and interpretation of large datasets. Based on SOM algorithm, one of the main visualization techniques is U-matrix, which is applied in this research.

In this research, text data for training the network is used. Thus, getting to a suitable neural network model with text compatibility was taken as the objective. Hopfield neural network and Back propagation neural network were built. However, based on text data specification, the SOM neural network was found to be very close to the objectives. The purpose was to identify a suitable neural network for text information retrieval. Therefore, a small scale network was designed. In large document collection (in practical mode) the following steps should be carried out for information retrieval:

1. Subject classification

In order to classify documents by subject, usually text automatic indexing systems are used, which assign text documents to subject groups. The benefit of this classification is that we have several groups with related documents that retrieve the relevant document very quickly. From these related groups and each query, we can access the related cluster, which will enable us to search directly. For this classification, we can use neural network. Since there is high dimensional property space of text documents, the neural network training with raw data in high dimensional space is quite time taking and slow. Therefore, it is proposed that property reduction techniques such as DF[3], CF-DF[4], TF x IDF[5] be used [13].

2. Text document specific recognition

Text document specific recognition embodies the following components:

-Word extraction
- Stopwords elimination
- Stemming
- Term weighting vector calculation
- Building suitable dictionary

3. Calculation of input property vector for each text document

4. Using the algorithm in Figure 1 to learn the SOM neural network

In attention to the neural network properties, it seems that applying AI techniques is very useful in information retrieval [3].

At present, the neural network model, are used a lot in information retrieval research. It is supposed that, in the future, with the progress observed in hardware and software, we can use neural network models for effective information retrieval [2]. According to hardware price reduction, it is possible to implement parallel document clustering with Neural Network by which a fast information retrieval is made possible [6].

With progress in software engineering, it is possible to have new methods for neural network control that will enable us to implement neural network models for high speed information retrieval algorithm.

ENDNOTES

1. SOM Toolbox for Matlab contains functions for creation, visualization and analysis of self organizing Map. The Toolbox can be used to preprocess data, initialize data and train SOMs using the range of different kinds of topologies. This toolbox is available on: http://www.cis.hut.fi/projects/somtoolbox.
2. U-matrix (unified distance matrix) representation of the Self-Organizing Map, visualizes the distances between the neurons. The distance between the adjacent neurons is calculated and presented with different colorings between the adjacent nodes. A dark coloring between the neurons corresponds to a large distance and, thus, a gap appears between the codebook values in the input space. A light coloring between the neurons signifies that the codebook vectors are close to each other iSn the input space. Light areas can be thought as clusters and dark areas as cluster separators. This can be a helpful presentation when one tries to find clusters in the input data without having any a priori information about the clusters.
3. In DF (Document Frequency) method, the categorization information is used to group the training documents such that all documents belonging to the same category

are put into the same group. After the documents are grouped, it is possible to form groups of the indexing terms in the vocabulary by putting, in a group, all terms contained in documents belonging to the same category. This process results in a set of sub-vocabularies corresponding to each category. For each document group, the document frequency is defined as the number of documents within that particular group containing the term. Therefore, important terms are those that appear frequently within a group of documents belonging to the same category. This is because the set of terms which are good representatives of the category topics should be used by most documents belonging to that category.

4. For any document group, a term appears in that group if at least one of the documents in that group contains that terms. For any term in the vocabulary of the training document set, the category frequency (CF) is equal to the number of groups that the term appears in. In this method, a threshold on the category frequencies of terms is defined such that a term is selected only if its CF is lower than the threshold, then the DF method is applied for further term selection to produce the reduced feature set.
5. A good measurement of the importance of a term in a document set is the product of the term frequency (TF) and the inverse document frequency (IDF), which is widely used in term-weighting based text retrieval system. The terms are ranked according to their TF x IDF values and the terms with the highest production values are selected to form the reduced feature set.

REFERENCES

[1] Crestani, F. A., "Model for Adaptive Information Retrieval." *Journal of Intelligent Information Systems*, Vol. 8, No. 1, 1997.

[2] Doszkocs, T. E., "Connectionist Models and Information Retrieval." *Annual Review of Information Science and Technology*, Vol 25, pp. 209-260, 1990.

[3] Fausett, L., *Fundamentals of Neural Networks: Architectures, Algorithms and Applications*. Prentice Hall, 1994.

[4] Grossman, D. A. and Frieder, O., *Information Retrieval: Algorithms and heuristics*. Kluwer Academic Publishers, 1998.

[5] Grunfeld, L., "Routing Retrieval and Filtering Experiments Using PIRCS." *Text Retrieval Conference*, U.S.A., 1995.

[6] Hatano, K., "A Som-Based Information Organizer for Text and Video Data." *Proceedings of the Fifth International Conference on Database System for Advanced Applications*, Australia, 1997.

[7] Kaski, S., "WEBSOM – Self-organizing Maps of Document Collections."

Neurocomputing, Vol. 21, Nos. 1-3, 1998.

[8] Koua, E. L., "Using Self-Organizing Maps for Information Visualization and Knowledge Discovery in Complex Geospatial Datasets." *Proceedings of the 21 International Cartographic Conference (ICC)*, South Africa, 2003.

[9] Laaksonen, J. T. and Koskela J. M., "Class Distributions on SOM Surfaces for Feature Extraction and Object Retrieval", *Neural Networks*, Vol. 17, Nos. 8-9, 2004.

[10] Lagus, K., "Mining Massive Document Collections by the WEBSOM Method." *Information Science*, Vol. 163, Nos. 1-3, 2004.

[11] Oja, M., "Bibliography of Self Organizing Map (SOM) Papers." Available: at <www. cis. hut.fi/research/som-bib>.

[12] Ultsc, A., "Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series." Available: at <http://citeseer.ist.psu.edu /ultsch99data.html>, 1999.

[13] Yin, L., "Learning Text Categorization by Backpropagation Neural Network." *Thesis*, 1996.