2004－ICS－138　(4)
2004／12／4

SBSOM: Self-Organizing Map for Visualizing Structure in the Time Series of Hot Topics

Ken-ichi Fukui,† Kazumi Saito,†† Masahiro Kimura††
and Masayuki Numao†††

In this paper, we propose a Sequence-Based Self-Organizing Map(SBSOM) that organizes clusters as series within the map to visualize their structure in terms of hotness, period and relations among topics. Principal Component Analysis(PCA) that is based on probabilistic document generation model is applied to extract hot topics from vast amount of documents, and these hot topics are used to label each document. Afterwhich, SBSOM is used to visualize these hot topics in a time series. SBSOM is also extended by defining label confidence for a more accurate labeling of its neurons. The initial experiments that use two kinds of news articles, the largest expands across ten years, validate that in addition to SOM showing only hotness of topics and relations among topics throughout whole period, SBSOM shows hotness within certain times, relations among topics, and period of topics.

1. INTRODUCTION

In recent years, vast amount of various documents in the WWW can be easily acquired. However, such mere amount of documents makes it difficult for users to be aware of much of their contents.

Documents that come out every day, especially news articles, can be naturally organized into topics. Furthermore, each topic can be characterized by:

Hotness: How many articles talked about the topic at a certain time?

Period: When the topic appeared and then disappeared?

Relation: Which topic is related to the topic?

To embody these characteristics so that one can instinctively grasp them in an entire picture, it is quite effective to use time series structures such as when a particular topic appears, increases in hotness then fades away, or is placed near similar topics.

Current visualizing approaches do not explicitly show all of these characteristics. Topic Detection and Tracking(TDT) is an approach that tries to find the solution for this lack of awareness by means of detecting specific topics automatically and tracking them from past to present among large number of time series information from multimedia such as newspaper, radio and television[1)2)]. In the TDT project using Multidimensional Scaling(MDS)[3)], though relations among documents can be seen, hotness and period does not explicitly appear within the map. Following the TDT study, TimeMine[4)5)] generates familiar timelines through its interface. Although it can handle hotness and period, it does not show relations among topics since it can only deal with a single topic.

Given these limitations, we focus on Self-Organizing Map(SOM)[6)] in order to handle all three characteristics. SOM is an unsupervised competitive artificial neural network learning that achieves both clustering based on similarities between input feature vectors and visualizing the clusters within a map at the same time. Therefore, SOM generates a map as taking into account relation among several topics and the number of documents corresponding to the topic. Although SOM provides an overview of document collection in text classification in terms of hotness and relation (e.g., WEBSOM[7)8)]), the map does not show period because SOM does not provide the time axis.

In this paper, we propose a Sequence-Based Self-Organizing Map(SBSOM) that presents a modification of the standard SOM. SBSOM provides the time axis to the map in addition to properties of SOM. Clusters in the map are labeled using hot topics. For hot topic extraction, Principal Component Analysis(PCA) that

† Dept. of Information Science and Technology, Osaka University
†† NTT Communication Science Laboratories
††† The Institute of Scientific and Industrial Research, Osaka University

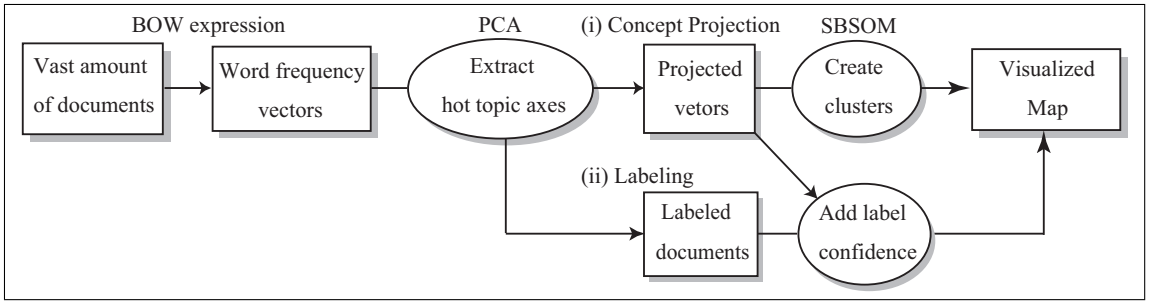

Fig. 1 Learning framework

is based on latent topic (naive bayes) model[10] is applied. Current SOMs use labels that were provided manually. However, in our method, labels are provided automatically by using extracted hot topics. At the same time, SBSOM extends SOM by defining label confidence for a more accurate labeling of its neurons because these labels are predictive.

The initial experiments that use two kinds of news articles, the largest expands across ten years, validate that in addition to SOM showing only hotness and relations among topics throughout whole period, SBSOM shows hotness within certain times, relations among topics, and period of topics.

2. LEARNING FRAMEWORK

Our learning framework is shown in **Fig. 1**. Each document is represented by word frequency vector using the Bag-of-Words model[11] which take into account occurrence frequency of a word that indicates how important it is for a given document. Afterwards, PCA obtains from the word frequency vector space axes representing the approximate content or notion of the hot topics.

The extracted hot topic axes are used in two ways:

(i) **Concept Projection:**
Word frequency vectors are projected into hot topic axis in order to decrease dimension by Concept Projection[12]. The frequency vector space becomes high dimensional with the increase in number of documents. Moreover, the word frequency vectors are sparse within high dimension space in most cases. Therefore, it is efficient to decrease dimension since it decreases computational time in SOM learning.

(ii) **Labeling:**
The other usage is to label each document with the hot topic whose axis coordinate within the projected vector has the maximum absolute value.

Then, SBSOM creates clusters based on document similarity within a map as time series structure. At the end, labeling by (ii) is used to visualize the map as clusters.

At the same time, label confidence is added for a more accurate labeling of its neurons. In our model, majority decision is adopted to select a class label for each neuron as a representative topic of the neuron. Since labeling by extracted hot topic axis is predictive, label confidence gives a weight that defines how much the labeled topic is suited to the document.

2.1 Using Standard SOM

Normally, SOM is constructed by a set of neurons which are arranged within a two dimensional grid with equivalent intervals. Our sample dataset documents are represented by $\{\mathbf{x}_n : n = 1, \cdots, N\}$, where $\mathbf{x}_n = (x_{n,1}, \cdots, x_{n,V})$, V represents the number of hot topics extracted by PCA, and $x_{n,i}$ is the i^{th} hot topic axis element given by Concept Projection for the document $\mathbf{x}_n$. Every neuron has a reference vector $\mathbf{m}_j$, which has the same dimension as the input data and corresponds to the j^{th} neuron($j = 1, \cdots, M$). When a input data $\mathbf{x}(t)$ is given by Concept Projection at the time t, SOM renew the reference vectors based on following equations.

$$\mathbf{m}_j(t+1) = \mathbf{m}_j(t) + h_{c(\mathbf{x}),j}(t)\,[\mathbf{x}(t) - \mathbf{m}_j(t)] \quad (1)$$

$$c(\mathbf{x}) = \arg\min_i \{\|\mathbf{x}(t) - \mathbf{m}_i(t)\|\}, \quad (2)$$

where the coefficient $h_{c(\mathbf{x}),j}(t)$ is called neigh-

borhood function which the first index $c(\mathbf{x})$ indicates the index of the winner neuron calculated by equation (2). To calculate distance between an input vector and a reference vector $\|\mathbf{x}(t) - \mathbf{m}_i(t)\|$, cosine similarity[13] is used. This is widely used as a measure for similarity between documents based on BOW text expression.

There are some definitions for $h_{c(\mathbf{x}),j}(t)$, but the one based on gaussian distribution is used in most cases:

$$h_{c(\mathbf{x}),j}(t) = \alpha(t) \exp\left(-\frac{\|\mathbf{r}_j - \mathbf{r}_{c(\mathbf{x})}\|^2}{2\sigma^2(t)}\right), \quad (3)$$

where $\mathbf{r}_j$ indicates the coordinate of the j^{th} neuron arranged within the two dimensional grid, $\alpha(t)$ and $\sigma^2(t)$ are parameters which control learning and use the strategy of reducing gradually and uniformly from certain value.

The original SOM is online learning that updates its parameters whenever each sample is given, as shown in equation (1) and (2). On the other hand, in this paper, offline (batch) learning that updates reference vectors after processing all samples is used since this improves learning efficiency[8].

2.2 SBSOM

We propose a Sequence-Based Self Organizing Map(SBSOM) by giving simple modification to the standard SOM. The standard SOM searches through all of the neurons for a winner neuron for each input data. SBSOM, however, restricts the search such that neurons which represent data that are in the same interval must be in the same column in the map. The input data are arranged in time series and are segmented in certain intervals. SBSOM then searches for a winner neuron from the same column (as shown in **Fig. 2**). Not only are data within the same interval placed in the same column, similar data, which have same or related topics, are placed nearly in the map horizontally owing to the neighborhood function.

Moreover, SBSOM gives meaning to the axis in the map in terms of time, while there is none in standard SOM.

2.3 Label Confidence

A label t_n of document $\mathbf{x}_n$ as its topic index is derived by selecting the maximum absolute value among its elements of input vector:

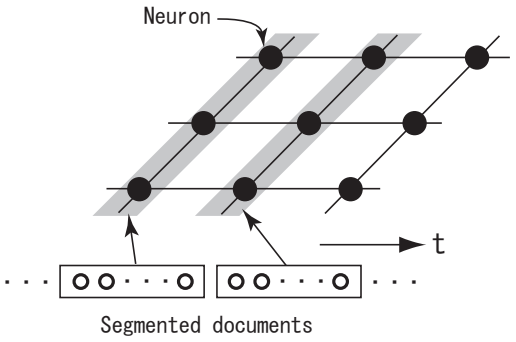

Fig. 2 Principle of SBSOM

$$t_n = \arg \max_{1 \le i \le V} |x_{n,i}| \quad (4)$$

After SBSOM is trained, the documents $\{\mathbf{x}_1, ..., \mathbf{x}_N\}$ are separated into classes $\{C_1, ..., C_M\}$, where C_j corresponds to the j^{th} neuron, and the j^{th} neuron label L_j which is representative topic of the neuron is derived by selecting the dominant class within the same neuron:

$$L_j = \arg \max_{1 \le i \le V} \#\{n : \mathbf{x}_n \in C_j, t_n = i\}, \quad (5)$$

where # represents the number of elements of a set.

Since our method adopts majority decision to select class label of each neuron, majority topic has the overwhelming advantage to be selected. Some topics which originally appear but minority may disappear due to this reason. This problem may be attributed to predictive labeling and the equivalent treatment made by majority decision on documents when class label of neuron is selected. Therefore, in order to help minimize this problem, we propose idea of label confidence indicated by means of weights.

The label for the following documents must have low confidence since there is a high possibility that they are mislabeled and should not be trusted. This implies that documents whose tendencies are opposite must have high confidence.

(i) A document containing small number of words that its vector is near to the axes origin.

(ii) A multi-topic document whose vector is near the 45 degree angle of the axis and the difference between first and second highest element is small.

Implying above tendency, the definition of label confidence is as follows.

$$Conf(\mathbf{x}_n) = \frac{|x_{n,T}||\overline{x}_{n,T}|}{\sum_{n=1}^{N}|x_{n,T}||\overline{x}_{n,T}|}N, \quad (6)$$

where

$$\overline{x}_{n,T} = \frac{x_{n,T}}{\sqrt{\sum_{n=1}^{N} x_{n,T}^2}} \quad (7)$$

and $T = t_n$ as the topic index of the document. The first term of the numerator in equation (6), $|x_{n,T}|$, solves (i), and the second term, $|\overline{x}_{n,T}|$, solves (ii). Clearly, $\sum_{n=1}^{N} Conf(\mathbf{x}_n)$ is equal to N, and that equation (6) means relative confidence.

2.4 Extension of Neuron Labeling

Using label confidence, the way of selecting neuron label is extended as follows:

$$L_j^{'} = \arg \max_{1 \le i \le V} \sum_{\{\mathbf{x}_n \in C_j, t_n = i\}} Conf(\mathbf{x}_n) \quad (8)$$

That is, select the class that the sum of label confidence is maximum within the same neuron.

3. EXTENSION OF MIP

MiP(Micro Averaged Precision)[14], which quantitatively measures the classification performance of SOM, is also extended when label confidence is introduced.

MiP evaluates correctness of neuron label as the dominant class in each C_j as the j^{th} neuron's class label. Assuming that y_n is the neuron label of the document $\mathbf{x}_n$, then standard MiP is defined as follows:

$$\text{MiP} = \frac{\sum_j \#\{n : \mathbf{x}_n \in C_j, t_n = L_j\}}{N} \quad (9)$$

MiP is extended by using label confidence as follows.

$$\text{MiP}^{'} = \frac{\sum_j \sum_{\{\mathbf{x}_n \in C_j, t_n = L_j^{'}\}} Conf(\mathbf{x}_n)}{N} \quad (10)$$

That is, use sum of confidence instead of counting the number of documents corresponding to neuron label. Note that these methods with label confidence naturally extend MiP and selection of class label because they give the same result as the standard way when all label confidence are set to 1.0.

4. EXPERIMENT

4.1 Experimental Condition

Two kinds of real world news articles dataset are used for experiments. One is the international news category of the famous Japanese newspaper "Mainichi", and the other is TDT2 corpus[1] that is used in the TDT project containing New York Times, CNN Headline News, ABC World News Tonight, and so on. These articles were collected from 1993 to 2002, and 1998, respectively. The number of articles and different words in each dataset are displayed in **Table 1**. Particularly, "Mainichi" in 1993 is dealt in detail in the experiment. The number of neurons is set to 24 in horizontal times 20 in vertical. We have carried out initial experiments that validate SBSOM's ability to embody the three characteristics as compared with SOM, as well as the efficiency of label confidence.

	M'93	M'93-'02	T'98
Articles	5,824	76,765	54,040
Words	24,661	72,155	113,898

Table 1 Basic statistics of the news articles dataset

4.2 Results and Evaluation

First, the comparison between maps generated by standard SOM and by SBSOM using the international news category of "Mainichi" in 1993 are shown in **Fig. 3**. Each symbol represents one hot topic, and major hot topics annotated manually are listed in **Table 2**. In the map using SBSOM (as shown in Fig. 3(b)), the horizontal axis indicates the time axis with monthly scale.

4.2.1 Hotness

Hotness of the topic is shown by the number of symbols corresponding to it in both maps. For instance, "Circumstance of Russia"(●), "Cambodian general election"(■), "Circumstance of China"(★), and "Bosnia conflict problem"(◎, ◆) are the hottest topics. From SBSOM, the number of the same symbol in vertical line indicates hotness of the topic at a certain month, meanwhile SOM shows it for the whole year. Furthermore, since number of each symbol tends to be similar between map(a) and (b), it can be considered that SBSOM preserves a property of SOM that the hotter topic occupies the larger area within the map.

4.2.2 Period

Contrary to map(a) where the period of the topic definitely cannot be seen, map(b) shows period of the topic as a horizontally continuous interval of the same symbol. The issue on "Cambodian general election"(■) gives a clear

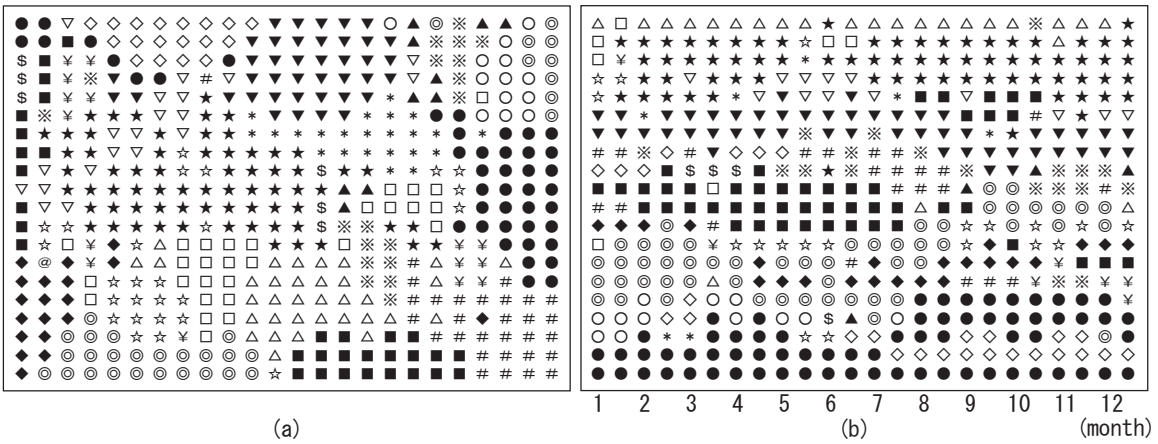

Fig. 3 Maps of the international news category of "Mainichi" in 1993 generated by; (a) SOM, (b) SBSOM.

symbol	hot topics
●	Circumstance of Russia -Opposing President Yeltsin and the national assembly -Creating a new constitution
◇	Middle East peace matter -Israel/PLO agreement and tendency of its related nations
■	Cambodian general election -Tracking from Cambodian general election to establishment of the coalition government
★	Circumstance of China -Economic circumstance -A change of government
◎	Bosnia conflict problem -Bosnia peace conference -Tendency of America and the United Nations
◆	Bosnia conflict problem -Tendency of Islam influence
▼	North Korea matter -Nuclear problem
△	Japan and foreign nations -Matter about compensation for wartime injuries or damage -Increasing its role in the United Nations

Table 2 Annotations of major hot topics

example. This appers until October in map(b), in fact, the general election was held on May 23rd and the new kingdom constitution was proclaimed on September 24th.

4.2.3 Relation

Both maps show relation among topics as similar topics are next to each other. For instance, (◎) and (◆) refer to the same topic "Bosnia conflict problem" but has different subtopics appear next to each other in both maps. Therefore, SBSOM preserves a property

	Hotness	Period	Relation
SOM	⋆⋆	-	⋆⋆
SBSOM	⋆⋆⋆	⋆⋆⋆	⋆⋆

Table 3 Comparison of evaluation by the three characteristics

	M'93	M'93-'02	T'98
A	73.8	82.2	72.6
B	55.3	70.6	59.2
C	70.4	92.3	68.4

Table 4 MiP′ value(%) (average of ten times)
A: SOM without label confidence
B: SBSOM without label confidence
C: SBSOM with label confidence

of SOM that similar topic which contains similar words frequency is placed near within the map. However, it cannot be asserted that topics next to each other are always similar topic. It depends on the relations among topics and on what topic was extracted by PCA.

4.2.4 Summary

To summarize above discussion, comparison between SOM and SBSOM for ability to embody the three characteristics are listed together in **Table 3**. Generally speaking, in addition to SOM showing only hotness and relations among topics throughout whole period, SBSOM shows hotness within certain times, relations among topics, and period of topics as preserving useful properties of SOM.

4.2.5 Efficiency of Label Confidence

Using three kinds of dataset, we evaluate the efficiency of label confidence as shown in **Table 4**. Since MiP′ becomes MIP when all con-

fidence are set to 1.0, MiP can be compared with MiP′. In all three dataset, MiP′ value was improved about 10 to 20%. Due to horizontal restriction in SBSOM, MiP′ value decrease around 15% compared with SOM. However, label confidence makes up for this loss.

5. CONCLUSION

In this paper, we have proposed Sequence-Based Self-Organizing Map(SBSOM) which embodies basic three characteristics of a topic, which are hotness, period, and relation by introducing the time axis within the map. The experiment using news articles validate that in addition to SOM showing only hotness and relations among topics throughout whole period, SBSOM shows hotness within certain times, relations among topics, and period of topics. Though current approaches do not explicitly show all of the three characteristics, SBSOM visualizes structure in the time series by embodying all of them at the same time. Furthermore, MiP, which is the quantitative evaluation measurement for SOM, was extended and improved by about 10 to 20% by introducing label confidence.

For application that makes use of SBSOM property, one immediate idea is to directly apply SBSOM in conjunction with news portal site for a search interface that will dynamically generate a map. Moreover, SBSOM can also be applied to any other time series data. We are now conducting experiments using hepatitis dataset for data mining by visualizing clusters of hepatitis patients depending on each feature, or by using results from SBSOM for preprocessing data for rule learning such as feature selection or heuristics acquisition of the data.

References

1) J. Allan, J.G. Carbonell, G. Doddington, J. Yamron, Y. Yang.: Topic Detection and Tracking Pilot Study: Final Report, *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Feb. (1998).
2) D. Frey, R. Gupta, V. Khandelwal, V. Lavrenko, A. Leuski, J. Allan.: Monitoring the News: a TDT demonstration system, *Proc. of the First International Conference on Human Language Technology Research*, (2001).
3) J. B. Kruskal and M. Wish.: Multidimensional Scaling, *Number 07-011 in Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park, CA*, (1978).
4) R. Swan and J. Allan.: Automatic Generation of Overview Timelines, *SIGIR-2000*, (2000).
5) R. Swan and D. Jensen.: TimeMines: Constructing Timelines with Statistical Models of Word Usage, *KDD-2000*, (2000).
6) T. Kohonen.: Self-Organizing Maps, *Springer-Verlag, Heidelberg*, (1995).
7) S. Kaski, T. Honkela, K. Lagus, and T. Kohonen.: WEBSOM – Self-Organizing Maps of Document Collections, *Neurocomputing*, 21:101–117, (1998).
8) T. Kohonen, S. Kaski, K. Lagus, J. Salojrvi, J. Honkela, V. Paatero, and A. Saarela.: Self organization of a massive document collection, *IEEE Transaction on Neural Networks*, 11(3):574–585, May (2000).
9) Saito K. and Ishikawa M.: Effects of Document Similarity Definitons on Classification Performance of Self-Organizing Maps, *Technical report of IEICE*, pp. 13–18,(2003).
10) Kimura M., Saito K. and Ueda N.: Extracting hot topics from the Web, *SIG-KBS-A401-21*, pp. 151–158, (2004).
11) Y. Seo and K. Sycara.: Text Clustering for Topic Detection, *tech. report CMU-RI-TR-04-03, Robotics Institute, Carnegie Mellon University*, January, (2004).
12) Sasaki. M, Otani. T, and Kita. K.: The Method of Constructing Concept Vectors for Information Retrieval, *IEICE NLP2002-32*, pp. 29-34,(2002).
13) Manning. C.D. and Schtze. H.: Foundations of statistical natural language processing, *MIT press.*, (1999).
14) N. Slonim, N. Friedman, and N. Tishby.: Unsupervised document classification using sequential information maximization, *Proc. of SIGIR*, pp. 129–136, (2002).