Using Self-Organizing Map for Ideas Clustering of Group Argumentation

Bin Luo and Xijin Tang
Institute of Systems Science, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing, China

Keywords: self-organizing map (SOM), idea clustering, visualization, and sum of square errors (SSE)

Abstract: Clustering may be used to refine a brainstorm into something that makes sense and give better organize information that may be dealt with more easily. This paper addresses self-organizing map (SOM) using for idea clustering during group argumentation process. We firstly introduce the principle and workflow of self-organizing map. And we design and implement an idea clustering tool based on self-organizing map (SOM) with Java. Thirdly, the clustering tool is used to analyze workshops of a famous scientific forum and we report the results and SSE measure of the clustering. Finally, we give some discussion and further research field.

1 Introduction

Brainstorm or daily group discussion produces ideas and opinions towards the concerned topic. Those ideas sometimes may be not well organized and in large number. There are lots of methods or tools that can help to group those pieces ideas in order to give better organized information. Tools like affinity diagram (also called KJ method) allow large number of ideas stemming from brainstorming to be sorted into groups for review and analysis. Qualitative meta-synthesis tools like CorMap and iView may help to conduct exploratory analysis and make sense toward the group discussion ideas (Tang, 2008; Tang, 2009). Technologies like text clustering and classification is a process that may extract effective, useful, understandable and valuable knowledge from large number of ideas and help to organize ideas. In this paper, self-organizing map is introduced and applied to idea clustering in order to get the main points of the group ideas and give something that makes sense to help understand group ideas in different perspective.

Self-organizing map is a very powerful and popular artificial neural tool used for a range of different purposes including clustering and visualization of high dimensional data spaces. The SOM algorithm has been applied to various areas such as speech recognition, robot arms control, optimization problems and analysis of semantic information (Lin, Soergel & Marchionini, 1991). It has been demonstrated that the SOM learning algorithm can perform relatively well in noise (Lippmarm, 1987) hence its application potential is enormous.SOM is said to “project and visualize high-dimensional data spaces” (Kohonen, 1990). The fact that there is a relation to clustering and visualization techniques is also well known. In the information analysis field, the clustering and visualization of SOM is also widely accepted. The self-organizing semantic map (Ritter & Kohonen, 1989) and Multi-SOM (Polanco, François & Lamirel, 2001) has been proposed and widely applied to cognitive information processing particularly offers the possibility “to create in an unsupervised process topographical representations of semantic, nonmetric relationships implicit in linguistic data” (Ritter & Kohonen, 1989). Following this vein, the SOM clustering and visualization for group discussion ideas is constructed.

The organization of this article is the following. Session 2 presents the detail and workflow of the Kohonen self-organizing maps (SOM). In Session 3, we show the implementation of the idea clustering tool based on self-organizing map. Session 4 discusses the experiment on some dedicated workshops of one topic in a famous scientific forum in China. And we report the results together with the measure of the SOM clustering for ideas. Finally, we report conclusion and future work to do in Session 5.

2 Self-organizing map

Self-organizing map (SOM) is a type of artificial neural network that is trained using unsupervised learning to visualize low-dimensional views of high-dimensional data. The architecture form of the SOM network is based on the understanding that the representation of data features might assume the form of a self-organizing feature map that is geometrically organized as a grid or lattice. SOM algorithm takes thus a set of N-dimensional objects as input and maps them onto nodes of a two-dimensional grid, resulting an orderly feature map. A layer of two-dimensional array of competitive output nodes is used to form the feature map. Every input is connected to every output node via a variable connection weight. Fig. 1 shows this network architecture.

Each unit in the gird or lattice is called a neuron, and adjacent neurons are connected to each other, which gives the clear topology of how the network fits itself to the input space. Input data sharing common characteristics will activate adjacent areas on the map. Each node in the grid is assigned an N-dimensional vector; the components of this vector are usually called weights. Initially weight components are small random values. Each input dimension is called a feature. An idea can be regarded as an input vector that is N-

dimensional vector. There are three essential processes involved in the formation of the self-organizing map as below.

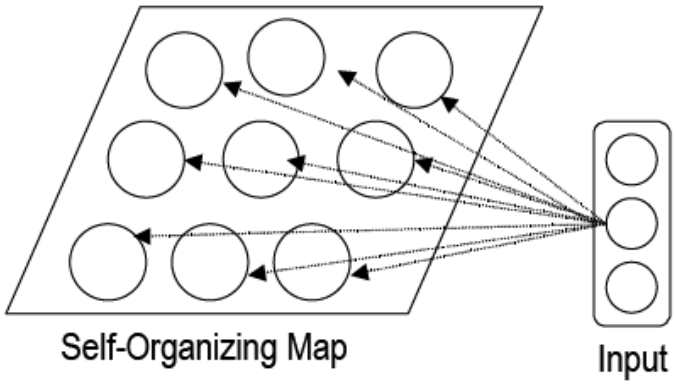

Fig. 1 SOM network architecture

Competition. An input vector should be selected randomly from the set of input vectors, and then the algorithm find the node whose weight closest to the input vector in the N-dimensional space (this node is called the winning neuron). The winning node is selected based on the Euclidian distance of an input vector and the weight vectors in the N-dimensional space. Let the $X(t) = [X_1(t), X_2(t), ..., X_N(t)]$ be the input vector selected at time *t* and let the synaptic weight vector of node j be denoted by $W^k(t) = [W^k{}_1(t), W^k{}_2(t), ..., W^k{}_N(t)]$ $k = 1, 2, \ldots l$ at time *t*, where *l* is the total number of neurons in the map. The winning node *s* is selected so that $\| X(t) - W^s(t) \| = \arg\min_k [X(t) - W^k(t)], k = 1, 2, \ldots l$. Briefly, the response of the process through the whole network could be either the index of the winning neuron, or the synaptic weight vector that is closest to the input vector in Euclidean sense.

Cooperation. The winning neuron determines the spatial location of a topological neighborhood of excited neurons, thereby providing the basis for cooperation among such neighboring neurons. Let h_{sj} denote a typical topological neighborhood centered on the winning neuron s, and j denote a typical neuron of a set of excited neurons around winning neuron s. Let the d_{sj} denote the lateral distance between winning neuron s and excited neurons j. A typical choice of h_{sj} that satisfies these two requirements is the Gaussian function

$$h_{j,s}(t) = e^{(-\frac{d_{j,s}^2}{2\sigma(t)^2})}, t = 0, 1, 2, ... \tag{1}$$

The lateral distance d_{sj} is defined as $d_{j,s}^2 = \| \overline{r_j} - \overline{r_s} \|^2$, where $\overline{r_j}$ and $\overline{r_s}$ defines the position of excited neuron j and s. A popular choice for $\sigma(t)$ which depends on discrete time n is the exponential decay described as below

$$\sigma(t) = \sigma_0 e^{\frac{t}{\tau_1}}, t = 0, 1, 2, ... \tag{2}$$

where σ_0 is the initial radius of the SOM algorithm and τ_1 is a time constant through the whole learning process.

Synaptic adaptation. After the winning node *s* is selected, the weights of *s* and the weights of the nodes in a defined neighborhood (h_{sj}) are adjusted so that similar input patterns are more likely to select this node again. By using discrete-time formalism, given the synaptic weight $W^k{}_j(t)$ of the neuron *k* at time t, the updated weight $W^k{}_j(t+1)$ at time t+1 is then defined by

$$W^k{}_j(t+1) = W^k{}_j(t) + \alpha(t) h_{k,s}(t)(X_j(t) - W^k{}_j(t)), 1 \le j \le N \tag{3}$$

The $\alpha(t)$ is the learning rate parameter, it should be time varying as indicated in Equation above for stochastic approximation. In particular, it should start at an initial value α_0, and then decrease gradually with increasing time *t*. It is shown by

$$\alpha(t) = \alpha_0 e^{(-\frac{t}{\tau_2})}, t = 0, 1, 2, ... \tag{4}$$

where τ_2 is another is time constant of the SOM algorithm.

Based on the three processes, the inputs and workflow of SOM algorithm is shown in Fig. 2. The lattice type of array can be defined to be square, rectangular, hexagonal, or even irregular. The most used forms are the square and the hexagonal arrays of nodes.

Note that the update procedure does not require any external "teaching" signals, so the algorithm is an unsupervised, self-organizing algorithm. Liked any unsupervised clustering method, SOM can find clusters from

the input data, and to identify an unknown data vector with one of the clusters. And the SOM represents the results of its clustering process in an ordered two-dimensional space. Such a mapping may effectively be used to visualize metric ordering relations of input data. In other words, SOM forms a semantic map where similar samples are mapped close together and dissimilar apart.

Input:

rs of the gird feature, and (x,y) is aware of each weight vector's location

$X \times Y$ is the number of weight vectors

σ_0 --initial neighborhood radius α_0 -- initial learning rate

t--current iteration n-- training times Wv--weight vector Iv-- input vector

Algorithm:

1) Randomize the map' nodes' weight vectors
2) *For t = 0 to n* then
3) Select an input vector randomly
4) Traverse each node in the grid
   Compute similarity between the input vector and weight vector
   Find the best match union (the winning node)
5) Update the nodes' weight vector in the neighborhood of BMU by Equation (3)
6) *end For*

Fig. 2 Inputs and workflow of SOM algorithm

3 Design and implement of self-organizing map

We design and implement an idea clustering tool based on self-organizing map. Fig. 3 shows the framework of clustering tool. The tool contain three parts—data preparation and pre-processing, SOM training and output and visualization.

Data preparation and pre-processing. The meta-data for the technology is of a structure as <*topic*, *speaker*, *speech*, *keywords*, *time*>. The *keywords* are articulated as attributes of *speakers* and *speeches*. A *speech* can be thought as an idea and also as an input for the SOM training. The meta-data can be organized in 4 type formats like XML, Excel, Access, and Txt. Speech-keyword frequency matrix is constructed based on the data records. Normalization should be carried out on the origin matrix and then the input vectors are formed. An input vector is a row of the normalization matrix.

SOM training. The input interface of the training process is shown in Fig. 4. The grid scale is determined by the input of X and Y dimension. Hexagonal or Rectangular can be selected as the lattice type. The neighborhood function and learning rate function also have different choice. The training time, initial radius and learning rate should be set up by users. And then the competition, cooperation and adaptation process will be executed automatically and weight vectors will be updated and output.

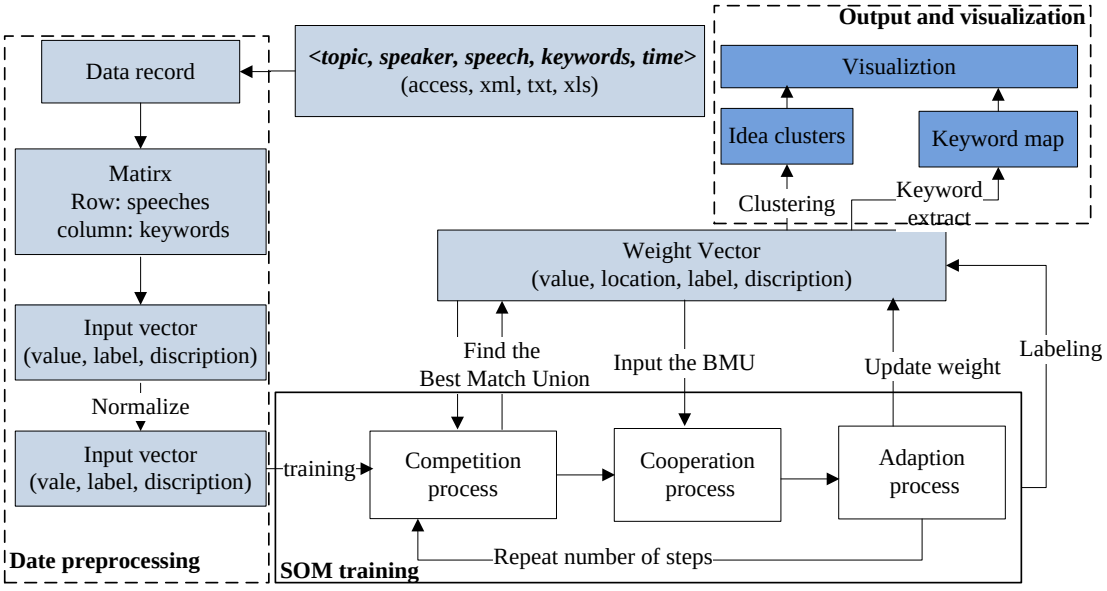

Fig.3 Framework of the idea clustering tool based on SOM

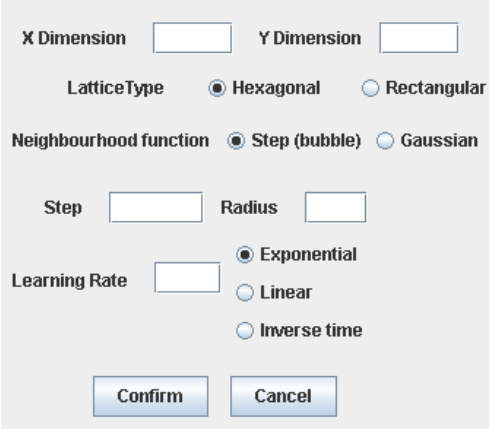

Fig.4 The input setting panel of the clustering tool

Output and visualization. After the training, the technology selects the keyword which index is the same as the index of the max weight value in on weight vector as the label of the node. Then the clusters are recognized by submerging the same labeled nodes and the labeled keywords are acquired as label of clusters. Finally, the clustering and visualization results are pushed to user.

4 Experiments: clustering of XSSC workshops

Xianshan Science Conference (XSSC) is a famous science forum which denotes a series of small-scale workshops which bring together a group of scientists working at the research frontier and enable them to discuss all aspects of the most recent advances in the field in depth and to stimulate new directions for research in China. XSSC has made important contributions to national science developments and exerted a profound impact on decision making process of the various government departments concerned. Then each workshop could be regarded as a group discussing process toward the concerned scientific problems (Tang & Liu 2006; Tang, Liu & Zhang, 2008).

4.1Dataset and preprocessing

In this experiment, all those conferences whose principal topics concentrate on "complex system and complexity research" are selected from over 300 conferences. 7 relevant workshops have been selected. A same data set analyzed by Ref. *Tang & Liu(2006) and Tang, Liu & Zhang(2008)* with minor modification towards keywords of speeches. For example, keyword like "systems science" in some speeches is amended to "systems engineering" according to the context of the speeches at the workshop. Then we totally get 23 speakers, 92 keywords and 61 speeches. One plenary speech in XSSC workshop can be represented as an input vector for the SOM training. For example, an item

<*Complex systems and complexity research, Yu Jingyuan, Complex Giant System Engineering, {open, giant system, complexity, qualitative, quantitative}, 1998-12-22*> indicates that scientist *Yu Jingyuan* gave a speech titled *Complex Giant System Engineering* about the topic *complex systems and complexity research* on Dec 22, 1998. O*pen, giant system, complexity, qualitative, quantitative* are 5 representative keywords of that speech. Here keywords of each speech are selected by the analysts and are regarded as the attributes or features of the speeches. Next, the clustering results by the SOM are given.

4.2 Clustering and visualization results

Fig. 5(a) is the result of SOM clustering under the input of initial learning rate=0.095, initial radius=4, 5×5 grid and training number=10000. 61 speeches (ideas) are sorted into 8 clusters (in different color) for further review and analysis by SOM clustering. Here those clusters are tagged as *"complex giant system"*(1), *"complex giant system and complexity"* (2), *"complexity and meta-synthesis"*(3), *"meta-synthesis"*(4), *"complex science"*(5), *"complexity"* (6), *"complexity and brain"*(7) *and "cybernetics"*(8) by human analysts. Both Cluster 1 & 2 may be combined into one cluster further accordingly. Those keywords may show the main point of the *"complex system and complexity"* discussion in XSSC workshop. According to the visualization results, the relationship of different ideas may be given by the SOM results. For example, the ideas labeled as "*complex giant system*" is close to ideas labeled as "*meta-synthesis*", and the same as "*complex science"* to "*cybernetics"*. Moreover, some innovation, novel or special ideas may be detected by the SOM algorithm. Such as the purple node tagged as 2 contains two ideas which are presented by *Song Jian* and *Qian Xuesen*. Here the analytical results are very good since the majority ideas are found (Cluster 1, 4, 5, and 8) together with important or special speeches are actually given by influential systems scientists in China. What's more, the visualization

results may help to understand and make sense of the ideas expediently. The number in the colored circle shows the number of the speeches that are mapped into the node. The details of the cluster will be displayed when the number in the node is clicked. Fig.5 (b) shows the speeches that are mapped into the dark blue node(Cluster 4).

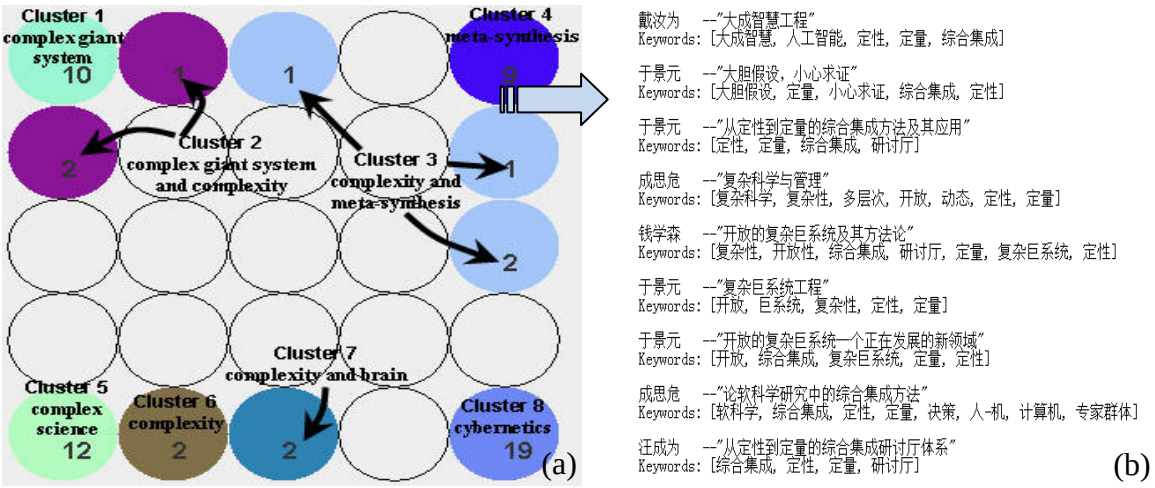

Fig.5 idea clustering of XSSC workshops by SOM

Frankly speaking, the clustering and visualization results are not always stable when the dataset is small scale because results are strongly correlated to the initial parameters. However, different results may give us different perspective of the original ideas. Moreover, the exploratory analysis given by SOM is useful to refine the group ideas into something that can make sense and give better organize information that may be understood and dealt with more easily.

4.3 Measure of the SOM clustering

The effectiveness of different clustering methods should be considered. SOM results may provide the representativeness keywords for each cluster and provide a probable perspective to understand the ideas. However, analysts can make further review and summary based on their understanding and experience. And different initial parameter should be tried to get a meaningful results.

Table 1 SEE measure of SOM and ADD for different dataset

Data set				ADD's SSE	SOM's SSE
Data set ID	Speech #	Speaker #	Keyword #		
I. workshops of year 1994 and 1995	11	9	27	10.74	4.60
II. workshops of year 94, 95 and 97	37	17	69	37.00	7.31
III. workshops of year 94, 95, 97,98,99	51	22	80	49.61	8.08
IV. workshops of year 94,94,97,98,99,02,04	61	23	92	57.89	8.62

*Initial parameter: ADD-row=4, column=4; SOM-X=Y=4, learning rate=0.095 radius=3, training times=6000

Generally, F-measure, entropy and sum of the square errors (SSE) are widely used measures of clustering results of the quality of clustering (Tan, Steinbach & Kumar, 2009).Both F-measure and entropy are supervised evaluation methods that need human experts' clustering results, which may not suit to the idea clustering because the group argumentation is always real-time process. However, the SSE may give the performance index of the SOM training. In this experiment, we organized the XSSC workshop into 4 different data set according to the year in which the workshop held. And we give a performance comparison of the SOM training and automatic affinity diagram (ADD) which is embedded in CorMap analysis (Tang, 2009). The SSE measure of SOM and automatic affinity diagram for 4 different dataset are given in Table 1. We find that for each dataset the SOM performs better than the ADD (with the lower SSE measure). What is more, with the number of ideas increasing the increasing rate of SSE measure is on the decrease. On the other hand, SSE may serve as a kind of factor that is related to the convergence of the group discussion.

5 Conclusion

In order to refine a brainstorm into something that makes sense and give better organize information that may be understood and dealt with more easily, we introduce the self-organizing map algorithm and demonstrate its potential use for idea clustering during the group discussion process. We design and implement idea

clustering tool based on the self-organizing map. The experiment based on the XSSC workshop data shows that the SOM may help to find the main point of the discussion together with some novel or special ideas. And also, the clustering and visualization results show the probable relationship of different ideas. On the other hand, the SOM used for idea clustering of group discussion is effectiveness and efficiency which is proved by the experiment. However, clustering is to group those pieces of ideas and the SOM training is just a kind of exploratory analysis and is strongly related to the initial parameter and then the role of human is more important. The explanation of clustering results is more interesting when depends on human's understanding.

Potentials of the self-organizing map are not limited to these clustering and visualization results. Further work along the main themes brought by the SOM will focus on design and implement of the interface for keyword clustering of ideas and keyword area map. Another important aspect is to experiment SOM to different data and try to find the fittest initial parameter for different context.

Acknowledgement

This work is supported by the National Basic Research Program of China (973 Program) under Grant No. 2010CB731405.

References

Kohonen, T., (1990), "Some practical aspects of selforganizing maps". *In IJCNN-90-WASH-DC, January 15-19, 1990, Volume II: Application Track*, Hillsdale, NJ, Lawrence Erbaum Associates, .p. 253-256.

Kohonen, T.,(1997), *Self-organizing maps*(*Second Extended Edition)*, Springer Series in Information Sciences 30, Springer-Verlag, Berlin Heidelberg.

Lin, X., Soergel, D., Marchionini, G., (1991), "A Self-Organizing Semantic Map for Information Retrieval", in *Proceedings of the 4th International SIGIR Conference on R&D in Information Retrieval*, 13-16 October, Chicago, p. 262-269.

Lippmrmn, R. P. (1987). "An introduction to computing with neural nets", *IEEE ASSP Magazine*, 1987(April), p.4-22.

Polanco, X., François C., Lamirel, J-Ch., (2001), Using artificial neural networks fo mapping science and technology : a multi-self-organizing maps approach, *Scientometrics*, 51 (1), p. 267-292.

Ritter, H. & Kohonen, T., (1989), "Self-organizing semantic maps", Biological Cybernetics, 61, p.241-254.

Tan, P.N., Steinbach, M., Kumar, V., (2006), *Introduction to Data Mining*, Pearson Addison-Wesley, Massachusetts.

Tang, X. J. & Liu, Y. J., (2006) , "Computerized Support for Qualitative Meta-synthesis as Perspective Development for Complex Problem Solving", in *Creativity and Innovation in Decision Making and Decision Support*, Decision Support Press, London, p. 432-448.

Tang, X. J., Liu, Y. J. & Zhang, W., (2008), "Augmented analytical exploitation of a scientific forum". *In S. Iwata, et al. (eds.) Communications and discoveries from multidisciplinary data*, Studies in Computational Intelligence 123, p.65-79.

Tang, X. J., (2008), "Approach to Detection of Community's Consensus and Interest", in *Proceedings of APWeb'2008 Workshops (Y. Ishikawa et al. eds.), LNCS 4977*, Springer-Verlag, Berlin Heidelberg.

Tang, X. J., (2009), "Qualitative Meta-synthesis Techniques for Analysis of Public Opinions for in-depth Study", *Complex 2009, Part II, LNICST 5*, Springer-Verlag, Berlin Heidelberg.