SOMSN: An Effective Self Organizing Map for Clustering of Social Networks

Fatemeh Ghaemmaghami
Research Scholar,
CSE and IT Dept.
Shiraz University, Shiraz, Iran

Reza Manouchehri Sarhadi
Research Scholar,
Dept. of Applied Computer Science,
UITM, Rzeszow, Poland,

ABSTRACT

Graph Clustering is a fundamental problem in many areas of research. The purpose of clustering is to organize people, objects, and events in different clusters in such a way that there exist a relatively strong degree of association between the members of each cluster and a relatively weak degree of association between members of different clusters.

In this paper, a new algorithm named self-organizing map for clustering social networks (SOMSN) is proposed for detecting such groups. SOMSN is based on self-organizing map neural network. In SOMSN, by adapting new weight-updating method, a social network is divided into different clusters according to the topological connection of each node. These clusters are the communities that mentioned above, in social networks. To show the effectiveness of the presented approach, SOMSN has been applied on several classic social networks with known number of communities and defined structure. The results of these experiments show that the clustering accuracy of SOMSN is superior compared to the traditional algorithms.

Keywords

Clustering, Social Network, Self-Organizing Map (SOM), Neural Networks

1. INTRODUCTION

A social network is a social structure consisting of individuals or organizations. This network can be considered as a mapping of all edges between vertices under study. Given the geographical spread of users, this kind of network is shown generally in the form of a social network diagram in which points are vertices and lines indicate edges that are relationship between the vertices.

Due to the high volume of data, social network analysis has become one of the favorite topics in data mining. An interesting problem in social network analysis is to identify communities in the network. Clustering is an important task for detection of community in networks. Finding existing patterns in data, by grouping individual and variables, is the main objective of this approach. The purpose of clustering is to organize people, objects, and events in different clusters in such a way that there exist a relatively strong degree of association between the members of each cluster and a relatively weak degree of association between members of different clusters. In other words, there should be low inter-cluster similarity but high intra-cluster similarity. Merriam-Webster defines the cluster analysis as "a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics" [1]. Community detection is important for some reasons. Identifying users who have similar interests may improve the efficiency of services on the World Wide Web [2] ,enables to set up recommendation systems with high performances [3], enables one to produce compact routing tables [4].

So far, a large number of clustering algorithms have been proposed for data classification based on centrality measures. Certain algorithms follow an iterative approach that begins operating by describing the whole network and or each of the nodes as society. When they begin with a whole network they use divisive methods [5], [6] and in case they begin with a node they use agglomerative methods [7].Freeman, in 1977, proposed vertex betweenness measure for identifying communities in social networks [8]. Girvan and Newman, in the years between 2002 and 2004, extended the definition of vertex betweenness to edge betweenness for identifying communities in social networks, and offered an algorithm for this purpose which has become one of the well-known algorithms in this regard [9]. Frey and Dueck, in 2007, developed a method called affinity propagation, which takes measures of similarity between pairs of data points as input [10]. CIBC is another algorithm for identifying communities in social networks. Compared with the detection algorithms that are used in different fields of science, CIBC has been designed to fulfill the job of detecting website communities from contents of a web server, in order to improve the performance of CDNs [11]. Zhao, *et al* proposes a new clustering method for community detection in which persons and their relationships among them are displayed by a weighted graph [12]. This method contains two remarkable features: 1) Creating a smaller hierarchical tree, which indicates current cluster.2) Recognition of overlapping clusters.

In this paper, a new algorithm for detecting communities in social networks is presented. This algorithm is designed based on self-organizing map (SOM). The advantage of using SOM is that the method can automatically (self-organizing) clusters nodes. The SOM algorithm also can be applied to a large scale of social network. The proposed algorithm uses a new weight assignment method in which a social network is divided into smaller sub-networks according to the communication between nodes of the neural network. In order to evaluate the proposed method, the algorithm will apply on well-known databases that have been utilized by many other individuals. For this purpose, the following databases are used: Karate Club, American College Football, and Books about US politics. The reason for using these databases that each represents a social network is because the structure, number, and members of each of these communities are clearly defined. Compared with other algorithms presented for identifying communities in social networks, this algorithm has a higher accuracy.

The rest of the paper is organized as follows: a brief review of SOM neural network is presented in Section2; the new clustering method is described in Section3; Section4 presents the applications of the method in social networks; finally, conclusion is drawn in Section5.

2. SELF-ORGANIZING MAP (SOM)

This is an unsupervised learning model. In this model, nodes or neurons are arranged in a two-dimensional space and the interaction between them defines the role of the SOM. This task is to estimate a distribution function [13].

In this method a number is selected for the number of output neurons. And using a simple logic calculates the geometrical distance of the model. Input and output neurons are initialized with binary values. The network works by reducing the distance between itself and the input patterns. Consider the vector X ϵ R^n where each of its elements has the probability density $P_i(x)$. Samples are selected periodically and randomly from this density space and applied on the network. According to the position of input vector in R^n the weights of the cells change according to algorithm. This change is done in such a way that ultimately weight vectors of the cells are evenly distributed in the probability density input space. Thus, the network estimates probability density of the input space by distributing its cells in it. The distribution of cells in the input probability space can be considered as a data compression. Because now, each cell is in the specified range represents an approximate of a specific range in space R^n [14].The Self-organizing map network algorithm is as follows (for a more analysis of the algorithm, refer to [15]) :

INPUT: A set of input vectors V = {V_1, V_2…, V_m}.
OUTPUT: A set of output vectors Z = {Z_1, Z_2…, Z_m}.
Step 1: Initialize weights by w_i^0. The neighborhood parameters of the winner are defined and the learning rate determined.
Step 2: While termination condition is false, repeat Step 3–Step 9.
Step 3: For each input vector X, repeat steps 4 to 6.
Step 4: Determine the neuron J as:

$$\mathrm{D}(j) = \sum_i \left(w_{ij} - x_i\right)^2 \qquad (1)$$

Step 5: Determine the index of J which has the closest distance to the input pattern (the min value of D (j)).
Step 6: Calculate the weight vectors of both the winner neuron and its neighbors as follows:

$$w_{ij}{}^{\text{new}} = (1-\alpha)w_{ij}{}^{\text{old}} + \alpha x_i \qquad (2)$$

Step 7: Update the learning rateα.
Step 8: Reduce neighborhood radius at specified times.
Step 9: Test the terminating condition.

3. A NOVEL CLUSTERING METHOD (SOMSN)

Consider a social network as a graph of n vertices. Adjacency matrix of the graph of the social network is shown by A = $[a_{ij}]$ n×n, which is a matrix that includes information about number of edges between different vertices of the graph. More precisely, the adjacency matrix of a finite graph G with n vertices is a matrix of n×n. the a_{ij} element is equal to the number of edges that connect two vertices V_i and V_j.

- Element a_{ij} is equal to 1 if the edge from V_i to V_j exists.
- Element a_{ij} is equal to 0 if the edge from Vi to V_j does not exist.

According to the above v_1, v_2... v_n will be placed in m output clusters c_1, c_2... c_n in clustering social networks by self-organizing map. Matrix W is considered as weight matrix and is shown as W = $[w_{ij}]_{n\times m}$ which is a matrix with n×m elements and the element W_{ij} defines the amount of influence of input V_i on output C_j and indicates that to what extent element Vi is related to cluster j. If input v_i is mapped to output C_j then C_j will be the winning neuron and weight of C_j must be updated according to equation (3). Then the corresponding weight vectors of all neurons within a certain neighborhood of the winning neuron are set according to equation (4). After applying the input V_i the weight vectors of the winning neuron and all neighboring neurons will move towards the vector C_j. After a number of iterations and providing different inputs to the network, neighboring neurons will learn similar vectors. The matrix Z is an n × m matrix, which is the output of the algorithm. If i belongs to cluster j, then Z_{ij} = 1 otherwise Z_{ij} = 0. After running the algorithm, n members of social network graph will be placed in m clusters which are stored in this matrix. Other algorithms such as K-Means, in order to cluster a social network, need to convert the graph of the network based on Euclidean or Hamming distance [16]. Here, using SOMSN, there is no need to change the input data, just it needs to apply neighborhood graph of the social networks as input to the algorithm that is used for applying the input in this method. According to the description given above, the proposed algorithm is as follows:

INPUT: Social network matrix shows the relationship between each node in the social network.
OUTPUT: An array of nodes which show specific cluster in the social network.
Step 1: The social network's adjacency matrix A and the matrix Z are created as defined above.
Step 2: Weights are initialized by w_i^0. The neighborhood parameters of the winner are defined and the learning rate determined.
Step 3: Repeat steps 4 to 10 while the termination condition is false.
Step 4: For each input vector X, repeat steps 5 to 8.
Step 5: Determine the neuron J as:

$$\mathrm{D}(j) = \sum_i \left(w_{ij} - x_i\right)^2 \qquad (1)$$

Step 6: Determine the index of J which has the closest distance to the input pattern (the min value of D (j)).
Step 7: For all units in the neighborhood of the winning neuron, the weight vector is set as follows:

$$(\mathrm{wij})^{new} = \frac{\mathrm{wij}^{old} + \alpha\mathrm{x} + \mathrm{Z}}{||\ \mathrm{wij}^{old} + \alpha\mathrm{x}\ ||} \qquad (3)$$

Step 8: For other non-neighboring units of the winning neuron, the weight vector is set as follows:

$$(\mathrm{wij})^{\mathrm{new}} = \frac{\mathrm{wij}^{\mathrm{old}} + \alpha(1-\mathrm{x}) + \mathrm{Z}}{||\ \mathrm{wij}^{\mathrm{old}} + \alpha(1-\mathrm{x}\ )\ ||} \qquad (4)$$

Step 9: Update the learning rateα.
Step 10: Neighborhood radius is reduced at specified times.
Step 11: Test the terminating condition.

4. EXPERIMENTAL RESULTS

In this section the performance of the new proposed technique compared to Girvan and Newman algorithm and the one presented by Zhao which are two of the most famous algorithms for detecting communities in social networks [12]. For this purpose, two well-known databases are used. The structure, the number of communities and their members clearly defined. The following benchmark data sets are selected for analysis:

1) Political books network
2) US college football

The properties of the above data sets are summarized in Table 1. In order to present the effectiveness of SOMSN algorithm in clustering social networks data, the clustering accuracy is measured by precision. Precision reflect the true number of nodes in the cluster that are relevant to the entire nodes in a social network. Precision calculate by equation (5) as bellow:

$$\text{Precision} = \frac{\text{Number of Corrected Nodes In Clusters}}{\text{Nuber of Nodes}} \qquad (5)$$

4.1 Datasets

4.1.1 Political Books Network Database

Books about US politics database was collected by Krebs in 2009 [9]. The database contains 441 records from 105 books. One of these books has been sold on Amazon. Subject of these books is the policies of the United States. In other words, the network has 105 vertices and 441 edges. A feature in Amazon shows that customers who have purchased the book have also bought other books. Based on this features, the edges between vertices in this network, indicate the relation between frequently bought books. Vertices in the network are labeled by letters L, N, and C representing liberal, neutral, and conservative respectively. This labeling of vertices has been done by Newman, in 2009, based on the description and reviews of the books posted on Amazon. There are three categories of books include liberal books, conservative books and neutral books. The goal is to distinguish these three categories of books.

Table 1: Summary of the properties of the real world data sets

Dataset	Number of clusters	Number of samples
Political books network	3	105
US college football	12	115

4.1.2 US College Football Network

Girvan and Newman introduced United States College Football network. This network was a representation of the schedule of football matches in the year 2000 [6]. In this network vertices represent teams and edges indicate a match between the two teams. There are 115 teams in this network, and 613 matches are held between different teams (see Figure 1).This is a network of known structure, and for this reason it has been used in many researches for detection of communities. The structure of this network is that teams are divided into conferences, each consisting of 8 to 12 teams.

The number of matches between teams within each conference is much more than the number of matches between teams from different conferences. Each team has an average of 4 matches played within the conference and 7 matches played with teams from other conferences. Intra-conference matches are not distributed evenly, and teams that are geographically close to each other are more likely to hold a match than the teams that are geographically at a further distance from each other.

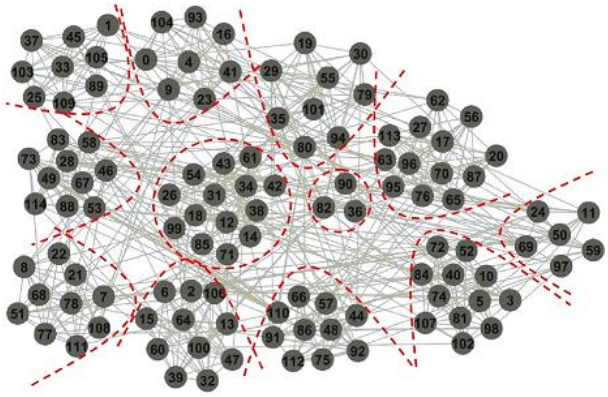

Fig 1. US college football network. In this network vertices represent teams and edges indicate a match between the two teams. There are 115 teams in this network, and 613 matches are held between different teams

4.2 Observation and Analysis

By applying Newman and Girvan algorithm on Political books database, the books are divided into three categories. Compared to the correct classification in this network, 17 books are not placed in their correct group. The number of books that have been incorrectly classified is as follows:(1, 5, 7, 8, 19, 29, 47, 49, 53, 59, 65, 66, 69, 77, 78, 104, 105).By Zhao's algorithm, the books are divided into three categories and again 17 books are not placed in their correct group.(1, 5, 7, 8, 19, 47, 49, 51, 53, 59, 65, 66, 68, 69, 77, 78, 86) (see Figure 2) [12].

By applying the proposed algorithm, the books are divided into three categories and the number of books that are incorrectly categorized is reduced to 16: (2-3-6-53-59 - 78-65-66-68-69-86-19-29-47-49-77) which shows an improvement compared to the two previously mentioned methods.

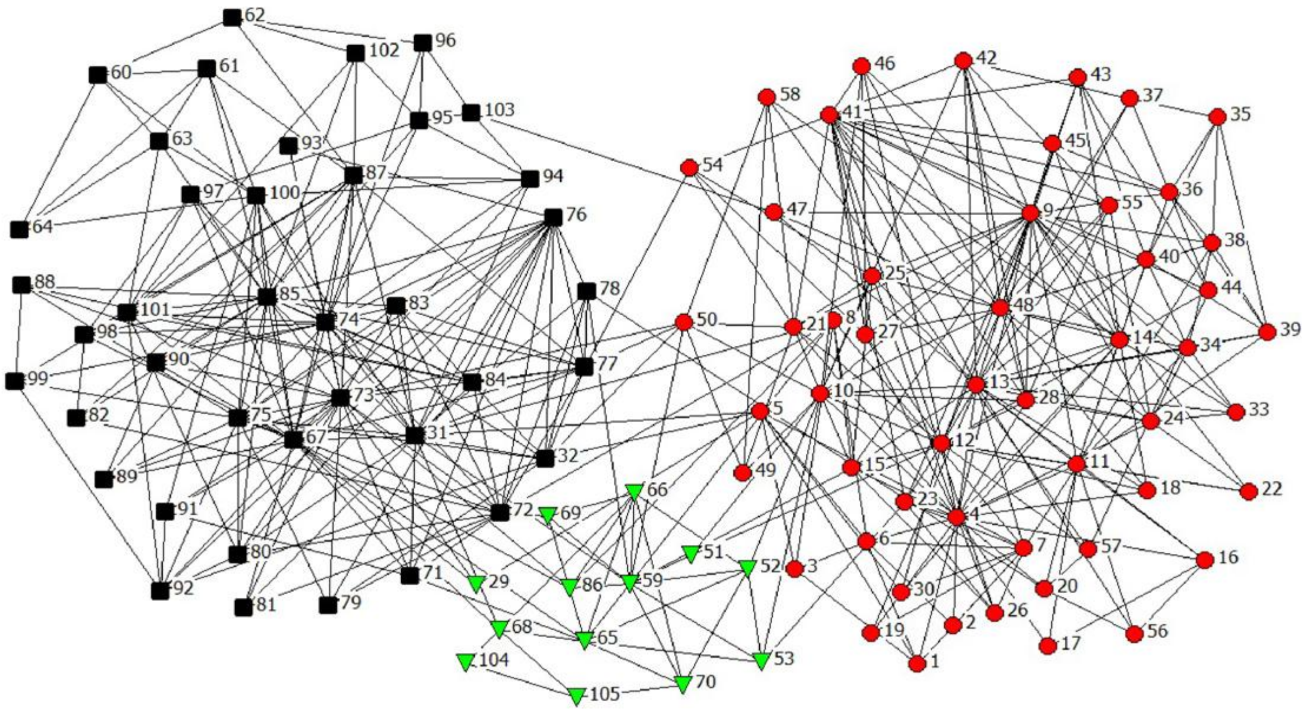

Fig 2. Result of Zhao method in political books, the books are divided into three categories [12]

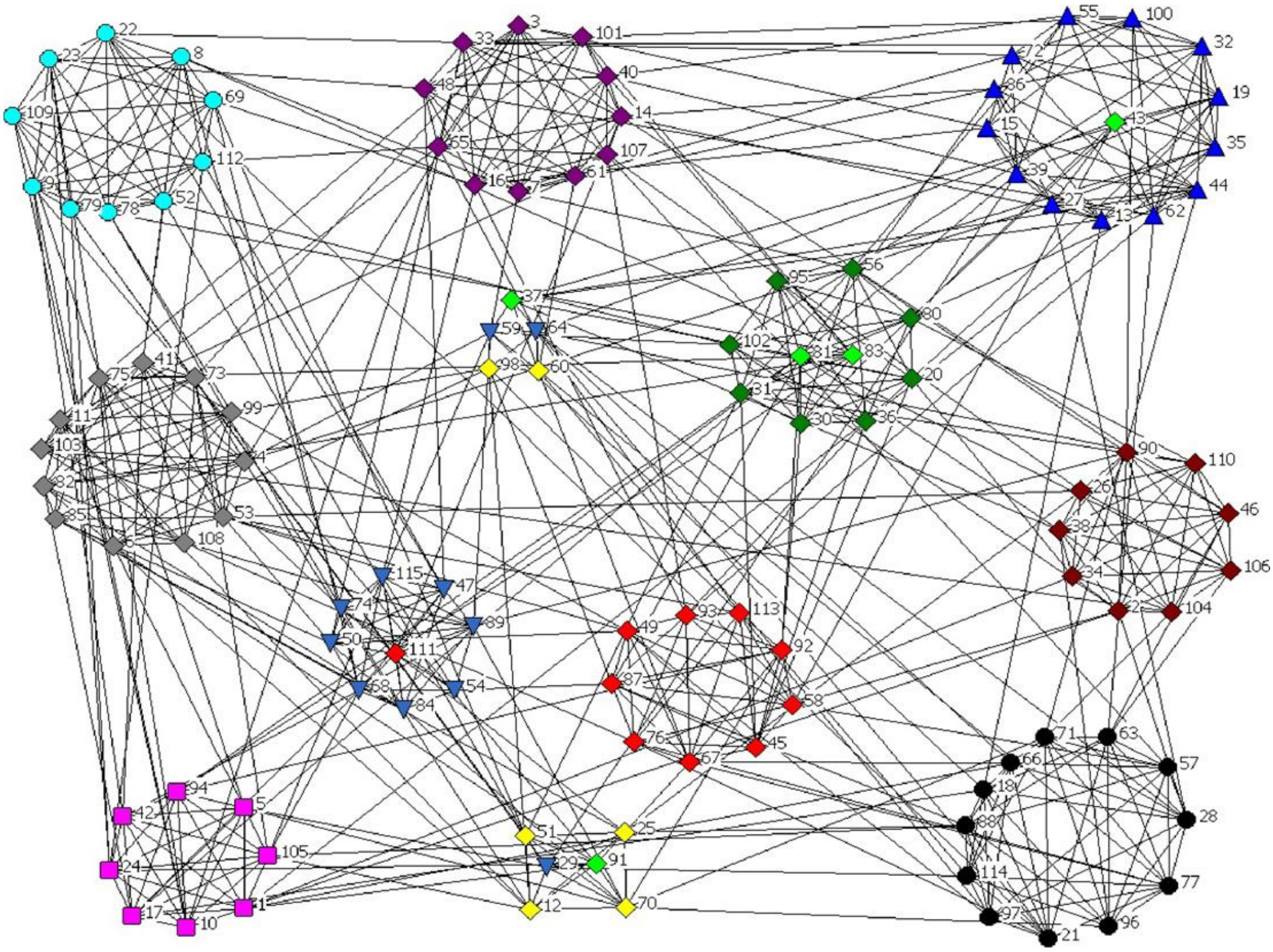

Fig 3. Result of Zhao algorithm in US College Football database [12]

By applying Girvan and Newman algorithm on the database, 12 conferences are formed and 11 teams are incorrectly placed in groups. By applying Zhao's algorithm on the database, 12 conferences are formed and 10 teams are incorrectly placed in groups (see Figure 3) [12]. By applying the proposed algorithm on the US College Football database, 12 conferences are formed and but only 9 teams are incorrectly placed in groups, which shows an improvement compared to the methods mentioned above. If, in this algorithm, the maximum number of clusters is incremented to 13 or 14 or even more, again, the number of conferences is 12, and the additional number of conferences will remain empty.

Table 2. Number of misclassified items for the social network data sets using the SOMSON and the Girvan and Newman algorithm and Zhao algorithm

Dataset	Girvan and Newman algorithm	Zhao algorithm	SOMSN algorithm
Political books	17	17	16
US college football	11	10	9

Table 2 shows a comparison between the algorithms described in this paper based on the number of vertices that have not been clustered correctly. A comparison between SOMSN algorithm, Girvan and Newman algorithm and the algorithm presented by Zhao, based on the Precision show the accuracy of the proposed method, In the case of benchmark data set, which used in this paper(see Figure 4).

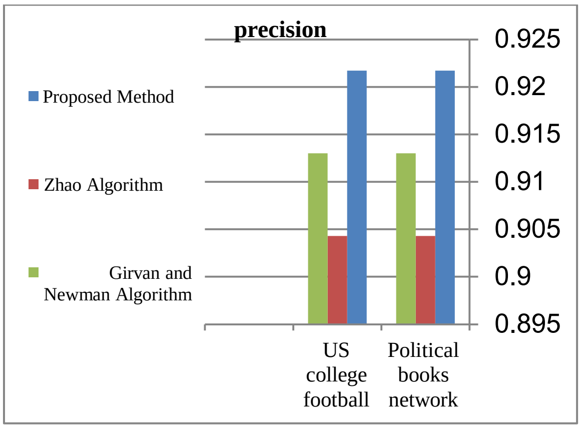

Fig 4. A comparison between SOMSN algorithm, Girvan and Newman algorithm and the algorithm presented by Zhao, based on Precision

5. CONCLUSION

In this paper, an algorithm is proposed based on self-organizing map (SOM). This algorithm accepts the structure of a social network as its input in the form of a neighborhood graph of the social network. By applying changes in the learning phase of SOM network, by adjusting the weight of neurons of the network, divides the social network into different clusters. Based on the results of applying this algorithm on various social networks, it can be observed that this algorithm is effectively capable of clustering a social network.

In future work, changes in other steps of the algorithm can be applied in addition to modification of the algorithm used for adjusting the weight of neurons, such as changing the method of calculating the winning neuron. Also the algorithm can be modified so that parameters such as density, network connections, etc are taken into consideration.

6. REFERENCES

[1] M. Webster, “Cluster analysis,” Merriam-Webster Online Dictionary, 2008. [Online]. Available: http://www.merriam-webster-online.com.

[2] B. Krishnamurthy and J. Wang, “On network-aware clustering of web clients,” ACM SIGCOMM Computer Communication Review, vol. 30, no. 4, pp. 97–110, 2000.

[3] P. K. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao, “A graph based approach to extract a neighborhood customer community for collaborative filtering,” in Databases in Networked Information Systems, 2002, pp. 188–200.

[4] N. Jeyaratnarajah, “Cluster-Based Networks,” in Ad hoc networking, 2001, pp. 75–138.

[5] B. Buter, N. Dijkshoorn, D. Modolo, Q. Nguyen, S. van Noort, B. van de Poel, A. Ali, and A. Salah, “Explorative visualization and analysis of a social network for arts: the case of deviantART,” Journal of Convergence Volume, vol. 2, no. 1, 2011.

[6] M. Girvan and M. E. Newman, “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences, vol. 99, no. 12, pp. 7821–7826, 2002.

[7] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, “Self-organization and identification of web communities,” Computer, vol. 35, no. 3, pp. 66–70, 2002.

[8] L. C. Freeman, “A set of measures of centrality based on betweenness,” Sociometry, vol. 40, pp. 35–41, 1977.

[9] M. E. Newman and M. Girvan, “Finding and evaluating community structure in networks,” Physical review E, vol. 69, no. 2, p. 026113, 2004.

[10] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” science, vol. 315, no. 5814, pp. 972–976, 2007.

[11] S. Papadopoulos, A. Skusa, A. Vakali, Y. Kompatsiaris, and N. Wagner, “Bridge bounding: A local approach for efficient community discovery in complex networks,” in arXiv preprint arXiv:0902.0871, 2009.

[12] P. Zhao and C.-Q. Zhang, “A new clustering method and its application in social networks,” Pattern Recognition Letters, vol. 32, no. 15, pp. 2109–2118, 2011.

[13] T. Kohonen, Self-organizing maps. Springer, 2001.

[14] T. Kohonen, “Self-organized formation of topologically correct feature maps,” Biological cybernetics, vol. 43, no. 1, pp. 59–69, 1982.

[15] D. Roussinov and H. Chen, “A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation,” Communication Cognition and Artificial Intelligence, vol. 15, no. 1, pp. 81–111, 1998.

[16] S. Haykin, Neural networks: a comprehensive foundation. NewYork: Prentice Hall PTR, 1994.