International Journal of Novel Computer Science and Power Solutions
(Available online at http://www.ijncsps.com)
April 2018 Volume 3. No.1
ISSN: 2476-7476

Presenting an optimal method for identifying communities in the Instagram's social network with clustering method

FatemePanabad[1], Seyyed Mohammad Safi[2*]

1- Department of Computer Engineering, Ahvaz Branch, Islamic Azad University Ahvaz, Iran
2- Department of Computer Engineering, Ahvaz Branch, Islamic Azad University Ahvaz, Iran
m.safi85@gmail.com

Abstract

In today's world, social networks play an essential role in expanding information.One of themost important issues is identification of communities in these networks.In previous methods, the identification of communities was such that the number of "Like" feedbacksfor each member in one group was determined regardless the number of member posts, and any member with more likes was identified as a more effective member.In the proposed method, the number of feedbacks to each member's post is determined by separating the posts of each member.As a result, the important posts are identified. Thus, the generated graph provides more comprehensive information about the communications of individuals. The new graph is then grouped by community identification methods such as Girvan Newman, CNM, Wakita Tsurumi and the final result will be presented as the final optimal result by voting on the results of the three above-mentioned methods.

Keywords: social networks, recognizing societies, community identification, community identification algorithms, clustering method

Paper submission date: October 2017

Paper accepted: December 2017

Paper Published: January 2018

Corresponding Author:

* **Seyyed Mohammad Safi**

Introduction

The growing use of communication in the virtual world and the increase in the number of social network users, the study and analysis of these networks are essential. In the social network, different people communicate with each other and people can share information, requests, advertisements and ... with each other.Each person in this network is represented as a node and the friendship relationship between them is shown as edges in the graph. One of interesting fields forsocial network analysts is the identification of communities in these networks. The purpose of community identification is to discover sub-structures that may exist in the networks. Identification of community is very important because the society is a small version of a complete graph that shows the very similar properties of that graph. So, surveying several communities may enable us to know how to use the entire network.This feature is very useful and important, especially when the network is a big data of the real world [1,12,13 and 14]. It is very important to recognize the structure of societies in the social networks. Social network analysis means studying and analyzing it.There are different methods for network analysis. These methods were originally originated from sociology and mathematics, or simply the graph theory. But today, in addition to these two disciplines, it is also used in other sciences. For this reason, we present a new method in this research for identifying the structure of communities, which has added a very important advantage to previous algorithms [2,18].

Related Works

The general meaning of society in social networks is aggregation of the nodes that can include users and …,in such a way that the members of this community have the most interaction with each other.Different algorithms are presented to identify societies, some of the most widely used algorithms are presented below [4, 3].

Anderson's algorithm: A clustering modular technique with the policy of roaming betweenthe vertices and the random selection of them that make up 1/2 of the communities; with this probability, it moves between vertices and goes to the end until the community is fully formedand able to improve the algorithm's execution speed, compared to other algorithms [5].

Fuzzy Detection Algorithms: This method calculates the correspondence power communities between all pairs of nodesand communities.In this algorithm model, a soft membership vector or membership coefficient for each node is calculated [6,19].

Bayesian Inference Algorithm: This is one of the statistical inference methods used in themodeling of real graphs, such as social networks.The Bayesian Inference uses observation to estimate the correctness of a hypothesis [8, 7].

Dynamic algorithms: This kind of algorithms work directly on the graph and extract the associations [7].

Methods of Nodexl

Nodexl is a software for analyzingsocial networks with many capabilities such as:

drawing theidentified groups and communities in the network, which can be done based on well-known algorithms. Some of the grouping methods of communities in Nodexl are as follows:

cluster method: It is a technique for clustering. In clustering, members within a cluster are very similar to each other, but the members of each cluster have less similarity to those in other clusters. Data similarities and differences are evaluated based on the magnitude of the characteristics and similarity criteria. The cluster method is divided into three algorithms, which will be explained separately [20,21]

Girvan Newman Method

The general philosophy of this method is to find intercommunity edges and to remove them. If we remove all the intercommunity edges, these communities will form separate elements.The general idea of this method is that the intercommunity edges with larger communication centrality are removed from the graph in an descending order.This method consists of 4 steps: 1. Calculation of the communication centrality of each edge 2. Removing the edge with the largest amount of communication centrality 3. Recalculation of the communication centrality of the edges 4. Return to the second stage.

remains, and all the nodes remain without an edge. This method had many time complexities, and most importantly, it could not discover overlapping communities. [8, 9,10 and16].

Clauset Newman Moor Method

These methods are based on the fact that the nodes of a community have common features, and these common features can be grouped together. Contrary to the dividing methods,we first consider the nodes in cumulative methods as unconnected and separate, and then connect them to each other in accordance to their common characteristics, so that communities are achieved. In this method, the intercommunity edges are removed and only the edges of the community remain.This method, which is a hierarchical clustering, starts with N unconnected and edgeless nodes and the edges are progressively added based on the reduction of similarity to the network. First, the most similar edges are added and then other similar groups are added to the community in order to gradually form a community [11,17].

Wakita Tsurumi Method

This method, which uses a stabilization ratio, involves a balanced growth among communities. This significantly reduces the complexity and according to their predictions, this method correctly handles networks with up to 10 million nodes [9, 10,and 15].

Proposed Method

In the proposed method, we define the social network as a compound graph, in

formed in a subgroup of member's posts, and the communications between members and posts form itsedges.The sub-graph nodes are created according to the number of likes for each post that

exceed the threshold value.If the two read vertices from onerowwere thesame, it means that the person has placed his post. Now for this post, a node is added and then the next lines that other people have liked this post are sequentially read and for people whose posts have 1/2 of the average likes, new nodes are added.Thus, by reading the entire graph, all nodes and the edges between them are created.The new graph is plotted.It is expected that this graph will be larger than the previous one, because it puts nodes for all posts of the person, and corresponding edges are drawn from the node-related post.The procedure is that first the information in the Excel file is copied in a text file and then is read.This file is named # World Cup (name of our datasets) for which, if tworead vertices of a rowwere the same, means that the person has placed his post; now for this post one node is put and then the next lines that other people have liked this post are sequentially read and a new node is added for people whose posts have the 1/2 of the average likes.Thus, by reading the entire graph, all nodes and the edges between them are created.Now we count the number of likes for each post and get the average number of likes: if the number of likes was 1/2 of the threshold, the post is considered and a new node is added. At the end, all the information is put in a new text file(by putting the necessary fields of the text file in the Excel file).

Now the new Excel file is entered in NodeXL and the new graph is drawn, The advantage of the presented method compared to previous methods is that the number of posts and the number of likes for each post can be easily recognized. So the effectiveness of a user post in the community is easily recognized.

Results and evaluation

Table1 shows the assessment of community recognition methods, in which the number of nodes and edges are determined. A model for assessing the effectiveness of the proposed method can be considered in two parts.

In the first part, advantages of this method which are not present in other standard methods can be mentioned. The most important advantage of this method is the effectiveness of nodes in the network.

The database used in this method is Instagram's Hashtag World Cup, which consist of 30122 nodes and 36423 edges.

In the second part, theoperationof proposed method is considered by other methods. For this issue, standard data from the Nodexl graphgallerywebsite with#WorldCupname is used. The data used are standard data that are used in almost all methods of community identification.

A good algorithm shouldoptimizelow and high nodes, because the number of nodes is very different in social networks, and in some cases the number is distributed to the entire population in the global network.To find out the community identification, the proposed method was compared with Girvan Newman, CNM, and Wakita Tsurumi methods in two stages andwith two different nodes and edges values. Then, by voting onthe results obtained from

three methods, the community is detected.

Table 1, The results of applying different methods to identify communities

Total data	Number of nodes	Number of edges	Wakita method	Girvan method	CNM method	Suggested method
# World Cup	100	98	0.423	0.570	0.583	0.591
# World Cup	1000	10520	0.465	0.599	0.621	0.632

In this plotting of the World Cup data with 100 nodes and 98 edges, the results were as follow: wakita method= 0.423, the Girvan method=0.570, CNM method=0.583 and proposed method= 0.591.In our proposed method, we had better communities than the three above-mentioned methods.

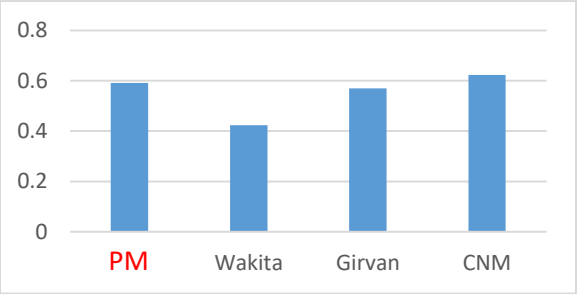

Figure 1, The results of community identification with 100 nodes and 98 edge

In the next stage, using the#World Cup total data with 1000 nodes and 10520 edges, the results were as follow: wakita method= 0.465, the Girvan method= 0.599, CNM method= 0.621 and proposed method= 0.632.In our proposed method, we had better communities than the three above-mentioned methods.

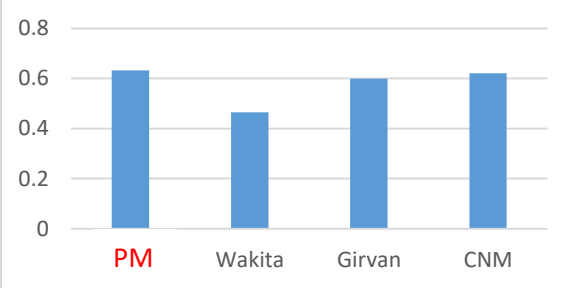

Figure 2, The results of community identification with 1000 nodes and 10520 edge

The assessment criterion of methods for true identification of communitiesis between -1 and +1. If the result is zero, then all the nodes are in one communityand if it is negative, it means that there are lots of mistakes infinding

communities.By reviewing all methods and also our method, it can be said that with any number of nodes and edges, the performance of the proposed method is better than other methods, and this improvement will be more optimal with higher number of nodes and edges.

Conclusion

A social network can include a number of nodes andthese nodes can include individuals or organizations and institutions,as well as a number of edges that indicate the relationship between them.In previous methods, all posts of each member were a node, and all the likes were entered in it. But, in the proposed method we consider a sub-graphfor each member,that corresponds to the number of themember's posts.Therefore, by observing the graph we can figure out the importance and impact of a post.For example, the difference between two nodes is easily recognizable when one of them has a low number of posts and a lot of likes, and the other has a lot of posts andless likes.

As a result, the number of feedbacks for each post is specified for each member.In this research, Girvan Newman, CNM, and Wakita Tsurumi methods are used for grouping communities in the Instagram social network.Finally,byreviewing all of the methods and the proposed method for identifying communities with any number of nodes and edges, it can be said that the performance of the proposed method is better than other methods. This improvement will be more optimal with the number of nodes and edges.Finally, the final result will be optimized by voting the results of the three methods mentioned above.

References

[1] P. Bródka, T. Filipowski, and P. Kazienko, "An Introduction to CommunityDetection in Multi-Layered Social Network," in Information Systems, ELearning,and Knowledge Management Research, Springer, 2013, pp. 185–190.

[2] SantoFortunato ,"Community detection in graphs", Physics Reports, Elsevier ,2010.

[3] Fortunato, S, " Community detection in graphs", Physics Reports Elsevier, No. 3, pp. 75-174, 2010.

[4] Tang, L. Wang, X. Liu .H, "Community detection via heterogeneous interaction analysis", Computer Science Data Mining and Knowledge Discovery, No. 12, pp. 1-33, 2012.

[5] Girvan, M. Newman, M, "Finding and EvaluatingCommunity Structure in networks", Physics Reports Elsevier, No. 69, pp. 26-113, 2004.

[6] Tang, L. Wang, X. Liu .H, "Community detection via heterogeneous interaction analysis", Computer Science Data Mining and Knowledge Discovery, No. 12, pp. 1-33, 2012.

[7] KAFHALI, S. HAQIQ, A. Liu .H, "Effect of Mobility and Traffic Models on the Energy Consumption in MANET Routing Protocols", International Journal of Soft Computing and Engineering, No. 1, pp. 2231-2307, 2013.

[8] Karsten, S. Nitesh, V, "Community Detection in a Large Real-World Social Network", University of Notre Dame IN USA, 2012.

[9] Xu, Y. Chen, L. Asaleh .S Nayak, R, "Network Detection on Metric Space", Ghrera Department of Computer Science and Engineering Jaypee University Of information Technology Waknaghat Solan Himachal India, No. 173215, 2015.

[10]. M. E. Newman and M. Girvan, "Finding and Evaluating Community Structure in Networks," Phys. Rev. E, vol. 69, no. 2, p. 026113, 2004.

[11] Fortunato, S, "Community detection in graphs", Elsevier Journal, 2010.

[12] Aggarwal, C. C. and C. K. Reddy (2013). Data clustering: algorithms and applications.

[13] Berkhin, P. (2006). A survey of clustering data mining techniques. In Grouping multidimensional data, pp. 25–71. Springer.

[14] Han, J., M. Kamber, and J. Pei (2011). Data mining: concepts and techniques: concepts and techniques. Elsevier.

[15] Wakita, K. and T. Tsurumi (2007). Finding community structure in mega-scale social networks:[extended abstract]. In Proceedings of the 16th international conference on World WideWeb, pp. 1275–1276.

[16] Girvan, M. and M. E. Newman (2002). Community structure in social and biological networks. Proceedings of the national academy of sciences 99(12), 7821–7826.

[17] Clauset, A., M. E. Newman, and C. Moore (2004). Finding community structure in very largenetworks. Physical review E 70(6), 066111.

[18] Asaleh, A, "Recommendation People in Social Networks Using Data Mining", Thesis submitted in fulfilment of the degree of Doctor of Philosophy, 2012.

[19] Nepusz, T., A. Petróczi, L. Négyessy, and F. Bazsó (2008). Fuzzy communities and the concept of bridgeness in complex networks.

[20] Strehl, A. and J. Ghosh (2003). Cluster ensembles—a knowledge reuse framework for combiningmultiple partitions. The Journal of Machine Learning Research 3, 583–617.

[21] Brandes, U., M. Gaertler, and D. Wagner (2003). Experiments on graph clustering algorithms.Springer.