ITM Web of Conferences **12**, 04019 (2017) DOI: 10.1051/itmconf/20171204019
ITA 2017

Social Relationship Discovery Via Call Records

Wen-Zhe ZHAO [1,a], Ping-Jian ZHANG [2,b]

[1]*School Of Software Engineering, South China University of Technology, Guangzhou, China*
[2]*School Of Software Engineering, South China University of Technology, Guangzhou, China*
[a]*392767725@qq.com,* [b]*pjzhang@scut.edu.cn,*

Abstract: Telecom users constitute a huge, but relatively sparse social network. Community discovery has been a research topic of data mining. Traditional algorithms are greatly influenced by outliers. This paper presents a new algorithm based on social triangle theory. Experiments show that the new algorithm is effective.

1 Introduction

Short messages and call records are an important part of social media network. One needs to dig out the various social relations among users or the behavior habit of the users from the call and SMS records, to develop more suitable pricing or business package politics for users, and to provide users with better service to attract more users.

To achieve the above objectives, the community partition of telecommunication users is a prospective direction. Before the birth of Internet, people began to study the structure of networked community, such as early biomedical research on the protein structure [1]. The division of community structure is closely related to the segmentation of images in computer science and the hierarchical clustering in sociology [2]. This paper utilizes the idea of social triangle theory and proposes a community discovery algorithm based on the improved triangle theory, and designs and conducts some experiments to demonstrate its effectiveness.

2 Related Work

2.1 Hierarchical Cluster Algorithm

Here we will introduce the splitting algorithm. The most representative splitting algorithm in the community algorithm is GN algorithm[3]. The basic idea of the GN algorithm is to find the largest edge in the network, and delete it. GN algorithm steps are as following:

Step1: find the betweenness of the entire network.

Step2: find the edge which has the highest betweenness, remove it.

Step3: repeat the step 2 until all notes are degenerated into a community.

The GN algorithm solves the problem that the Laplace bisection algorithm must know the number of nodes in the community in advance. In order to get better clustering, Newman et al. proposed a standard for measuring the quality of community division-modularity. Modularity is used to represent the ratio of the number of edges connecting two different communities to all edges in the network. The specific formula of the modularity is shown as follow:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j) \tag{1}$$

where m represents the sum of the number of edges in the network. A_{ij} represents the value of the i-th row and the j-th column element in the adjacency matrix. The matrix P is an adjacency matrix used to store the relationship between the original nodes, and P is the node i and j correspond to the value of the element ij. If the two nodes belong to the same community, ie $c_i = c_j$, then $\delta(ci, cj) = 1$, otherwise $\delta(ci, cj) = 0$.

In the pseudo-random network mentioned above, when the nodes i, j point to each other, said node i, j connected. Assuming that the degree of the two nodes in the network (including the out-degree and in-degree) were k_i and k_j. For the network with m sides, we can define $p_i = \frac{k_i}{2m}$ as the connection probability of i to j. $p_j = \frac{k_i}{2m}$ as the connection probability of j to i. Define any element in P is $P_{ij} = 2mp_ip_j = \frac{k_ik_j}{2m}$, then the Q can be rewritten as following:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{k_ik_j}{2m})\delta(C_i, C_j) \tag{2}$$

According to the modularity theory, we can see that the closer the Q is to 1, the more obvious the community structure is, and the Q found in the actual network community is often located between 0.3 and 0.7 [4].

3 Community Discovery Alogorithm Based On Improved Triangle Theory

3.1 Triangle Theory

In real society, the relationship among people is complicated, for example, A and B are good friends, B and C are also good friends, so are C and A. Thus, the relationship among them constitutes a triangle shown as below.

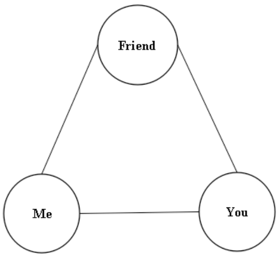

Figure 1. Simple social triangular structure diagram.

For members of the same social triangle, we can classify them as a unified community. Since there is such a social triangular relationship exists in the network, the network diagram given in the above can be extended to build a complex display society, as shown in Figure 2.

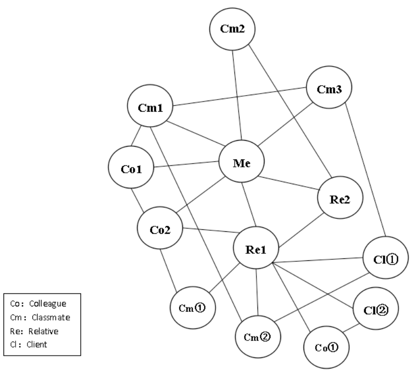

Figure 2. Complex social triangle group structure diagram

4 Design Of Triangle Community Division Algorithm

4.1 The Initial Community To Determine

We first determine the threshold ε of the good acquaintance, and then find out all the links and the similarity between nodes. Points with similarities larger than ε are grouped into a triangular group as the initial community.

4.2 The Algorithm Flow

Algorithm 2: Community Discovery Algorithm Based on Triangle Theory

Input: The structure of social network (diagram)
Output: Each community and the members of community
Steps:
//Step1: Find the initial triangular group.
Set ε = value;
for every node v_k in V **do**
  if v_k can make a triangle with other 2 nodes (v_i, v_j) (or more nodes)
  then
    generate new cluster_ID;
    insert v_k, v_i, v_j ……into queue Q_k
  else insert v_k into queue Q_{res}
  delete v_k form V;
until every node was assigned.
//Step2: Traverse all the initial societies and merge the initial communities containing the public triangles;
for j1=0 to k-1 **do**
  for j2=j1+1 to k
    if count $(v(Qj1 \cap Qj2)) >= 3$
      Merge (Q_{j1}, Q_{j2})
//Step3: To find structure-accessible point which is not distributed based on the new community composition obtained by step2
for every node v in Q_k do
  for every node u in Q_{res};
    if max($v \rightarrow u$)
      assign u to Q_k which most similar to u;
  delete u form Q_{res};

5 Experimental Design and Result

5.1 Experiment Result and Analysis

The algorithm designed in this paper is based on the clustering theory. The calculation process of this paper is actually a complex network clustering process, in the process of trying to find the optimal clustering results for the known clustering structure data We often use the accuracy rate, recall rate and F value to evaluate the effect of clustering [6].

Table 1. Accuracy and recall rate analysis table

	Assigned correctly	Assigned wrongly
Detected	T1	F1
Undetected	T2	F2

1) The first step is to perform an initial triangular search in all nodes. In the second step, the upper triangular group is merged and the members of the triangle are expanded by searching for the structural community as the initial community. Finally, three societies along with two outliers are detected. Since one of the three societies contains only three nodes, the three nodes are treated as incorrectly assigned nodes when calculating the accuracy rate. In this test, the similarity threshold is chosen to be 0.1. The final result is shown in table 2.

Table 2. clustering based on triangle improved algorithm result

Clustering Based on Triangle Improved Algorithm						
category	pi	pj	pi∩pj	Accuracy rate	Recall rate	F value
Associa-tio n1	15	16	15	1	0.938	0.968
Associa-tio n2	14	18	14	1	0.778	0.875
The average				1	0.858	0.922

Where pi represents the number of algorithm allocations and pj represents the number of original network members, pi∩pj represents the number of members of the algorithm that are overlapped with the original association.

The GN algorithm result is shown in table 3.

Table 3. GN algorithm clustering result

Clustering GN algorithm clustering						
category	pi	pj	pi∩pj	Accuracy rate	Recall rate	F value
Associa-t ion1	16	16	16	1	1	1
Associa-t ion2	13	18	13	1	0.722	0.839
The average				1	0.861	0.92

In the US political book network dataset, The similarity threshold is chosen to be 0.3 in this test. In this experiment, the algorithm proposed in this paper divides the whole network into five communities, and the GN algorithm divides into six communities, where the nodes with lower coverage are regarded as outliers. The specific experimental results are shown in Table 4 and Table 5.

Table 4. Clustering based on triangle improved algorithm result

Clustering Based on Triangle Improved Algorithm						
category	p_i	p_j	pi ∩ pj	Accuracy rate	Recall rate	F value
Associ-ation1	8	13	14	0.500	0.308	0.381
Associ-ation2	40	49	35	0.875	0.714	0.787
Associ-ation3	36	43	32	0.889	0.744	0.810
The average				0.755	0.589	0.659

Table 5. GN algorithm clustering result

Clustering GN algorithm clustering result						
category	pi	pj	pi∩pj	Accuracy rate	Recall rate	F value
Association1	7	13	3	0.429	0.231	0.300
Association2	45	49	41	0.911	0.837	0.872
Association3	42	43	35	0.833	0.814	0.824
The average				0.724	0.627	0.665

The results are depicted as follows:

Figure 1. Comparison of Different Parameter of Triangle Sociology and GN Algorithm in Two Data Sets

Combining the above results with GN algorithm, it can be seen that the algorithm designed in this paper can reach the level of traditional excellent community division algorithm.

6 The Application Of Triangle Algorithm In Telecommunication Network Division

In the experiment, about 20,000 call records among 524 users are extracted. The triangular community division algorithm and the GN algorithm are compared and analyzed respectively. The Dunn index is picked to evaluate the effect of clustering[5], whose expression is as formula.

$$D = \min_{i=1,\ldots,g}\{\min_{j=i+1,\ldots,g}(\frac{d(C_i,C_j)}{\max_{k=1,\ldots,g}(diam(C_k))})\} \quad (3)$$

Where $d(C_i,C_j)=\min_{x\in C_i,y\in C_j}\{d(x,y)\}$ represents the distance between two different clusters, $diam(C_i)=\max_{x,y\in C_i}\{d(x,y)\}$ indicates the maximum distance between members of the cluster. The larger the D is, the larger the cluster is. The results are shown in Figures 7.

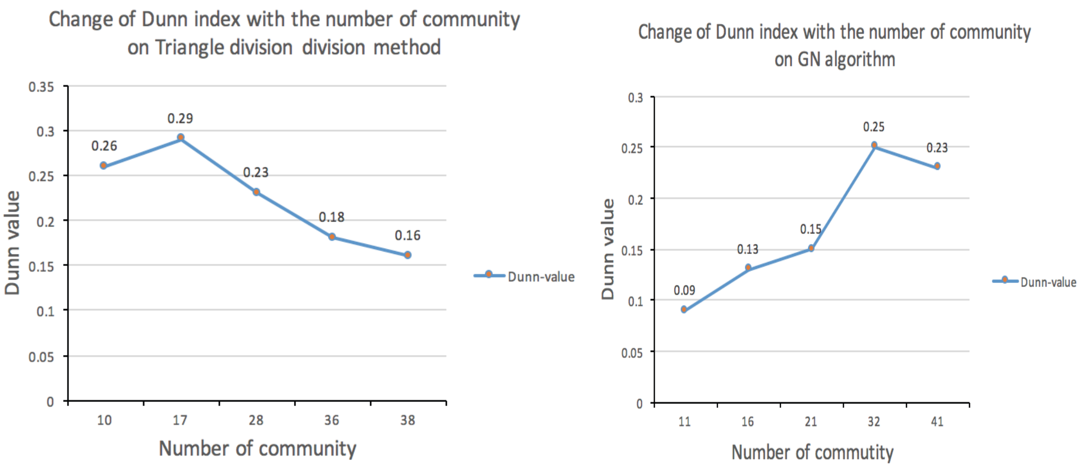

Figure 2. Diffrent method Dunn index changes with the number of community

The maximum value of Dunn index is 0.29, which is larger than the maximum value of Dunn index of traditional GN algorithm. This indicates that the triangular community division algorithm is better than GN algorithm when the number of community is kept at a certain value.

As a traditional community network division algorithm, GN algorithm for outlier detection and the capacity of processing telecom networks is relatively weak. The triangulation division algorithm is mainly based on the similarity of triangular relationships among node. If a user can not build a triangular relationship with other users, none of the structural reachable point of the user appears in a triangular group. The algorithm will treat it as an outlier. Due to the exclusion of a large number of outliers, the maximum value of the Dunn index of the triangulation method is higher than that of the GN algorithm, and the clustering efficiency is better than the GN algorithm.

Summary

In the paper, existing community discovery algorithms are reviewed. Then, a triangle community division algorithm based the social triangular theory is proposed and implemented, and validated using standard data set. The new algorithm is then compared to the GN algorithm using the telecommunication call record data set. Experimental results show that the new is algorithm is more effective.

Acknowledgement

This work is supported by the Guangzhou Science Technology and Innovation Commission (Grant No. 201604010099).

References

1. Li Haiyan. Study on Protein Folding with Complex Networks [D]. Beijing: Beijing University of Posts and Telecommunications, 2009, 18-21.
2. Girvan M, Newman M E J.Community structure in social and biological networks [J]. Proc Natl Acad Sci USA, 2002, 99:7821-7826.
3. ZHENG Feng-ni.Study on discovery method of network community based on node similarity clustering in complex network [J] .Computer & Modernization, 2013,5: 231-233.
4. PAN Gao-feng, WANG Xing-hua et al.Application of community discovery method in complex network partition identification [J]. Power System Protection and Control, 2013,41 (13): 116.
5. LUO Zhi-gang, DING Fan, et al.New Advances in Research on Complex Network Community Discovery Algorithm [J].Journal of National University of Defense Technology, 2011,33 (1): 47-51.
6. Fan Ming, Meng Xiaofeng. Concept and Technology of Data Mining [M]. Machinery Industry Press, 2012.7.
7. Pei Weidong, Xia Wei, et al.Analysis of Evolutionary Model for a Class of Triangular Structure Dynamic Complex Networks [J]. Journal of University of Science and Technology of China, 10 (11), 2010.
8. Lancichinetti A, Fortunato S, Kertesz J. Detecting the overlapping and hierarchical community structure of complex networks[EB/OL].2009.
9. Lü Tianyang, Zheng Weimin et al. Evaluation Index of Weighted Complex Network Society and Its Discovery Algorithm [J], Acta Physica Sinica, 2012,61 (21).
10. Statistical.Pattern.Recognition, Andrew.R.Webb, Keith.D..Copsey, 3ed, Wiley, 2011: 543-545.