Mining Telecommunication Circles via the Call Record and Short Messages

Shuai ZOU, Ping-Jian ZhANG

School of Software Engineering, South China University of Technology, Guangzhou, China
767186977@qq.com,pjzhang@scut.edu.cn

Abstract—Telecommunication circles are groups of similar customers in telecommunication networks. Mining such circles provides with telecommunication operators great value in developing prospective customers while retaining old ones. However, most of the existing community detecting algorithms utilize mainly the structure of the complex network and ignore the strength of relationship. This paper improves the classic CPM (Clique Percolation Method) algorithm by taking into account both the call record and short messages, and proposes a new algorithms called SR_CPM (Strengthened Relationship CPM). The new algorithm is applied to telecommunication networks and demonstrates superior effectiveness over CPM.

1 Introduction

Telecommunication users constitute a huge, but relatively sparse social network. Community discovery has become a very hot research topic of data mining for many years. Telecommunication circles are groups of similar customers in telecommunication networks. Finding out these circles is of great value for operators, helping them make more attractive pricing package politics, obtain prospective customers while retaining old ones. However, current community discovery algorithms in the literature mainly focuse on the people's personal behaviors as well as population attributes. This paper will employ information such as call record and short messages, and present an improved CPM algorithm called SR_CPM (Strengthened Relationship CPM). A group of experiments are designed and results show that the new algorithm is effective and efficient.

In this paper, part two introduce the related work, part three introduce an improved CPM algorithm and part four introduce the application of SR_CPM in telecommunication data set.

2 Related work

The complex networks have the characteristics of "small world", "scale-free", and community structure. Many social networks in real life can be abstracted as complex networks, and telecommunication networks are the typical ones. Farrahi[2] et al. have studied the location driven daily behavior patterns using classification method. Kim[10] et al. set up a logistic regression model to handle the Korea Telecom Data. Especially, community discovery technology is well suited to find out the community structure in complex networks. Early community discovery algorithms are derived from graph theory like the spectral bisection community discovery[3], or derived from the hierarchical clustering algorithm, ,which introduces a measure Q[4], to reflect the degree of modularity of community division. Based on this concept, Girvan and Newman proposed the GN algorithm[1] by deleting the edge of the network in the largest number of edges to the community Division. In addition to the non-overlapping community detection algorithm mentioned above, research workers also put forward a series of overlapping community discovery algorithms that employee the idea of clique filtering[5] , seed expansion[6] , hybrid probability model[7] and edge detection[8] , etc.

3 An improved CPM algorithm

3.1 Existing Problems of Community Discovery Algorithm

Although many community detection algorithms has been proposed, it is still a challenging work to find out community structure from the complex network. There are still a lot of problems need to be solved.

1. Most of the current algorithms are based on static networks.

2. The effect and performance of the community discovery algorithms have been a problem. It is a great

challenge to design highly effective algorithms with low time complexity.

3. Most of the existing community discovery algorithms only consider the connections between nodes while ignore the strength of the connection and the inherent attributes of the nodes.

3.2 Improved Clique Fitering Community Discovery Algorithm

The main contribution of this paper is to improve the CPM algorithm, called SR_CPM, which is more suitable for the community discovery in the telecom user call network.

3.2.1 Algorithm idea

There are mainly two kinds of methods to expand the CPM to weighted network.

a) Set a global threshold w, remove edges with weights less than w and then use the traditional CPM algorithm to partition the network to communities.

b) Farkas[9] et al. proposed a clique intensity function for CPM in the clique algorithm. For a k-clique containing k* (k-1) /2 edge, its clique intensity is defined as:

$$I(C)=\left(\prod_{\substack{i<j\\ i,j\in C}} w_{ij}\right)^{2/k(k-1)} \qquad (1)$$

And cliques whose intensity is less than some prescribed value are ignored.

The first method is simple but the choice of threshold w is difficult and greatly influences the quality of community partition. The second method is generally more effective but suffers from huge computations. Moreover, the intensity function is an absolute index hence and it is not easy to set the threshold either. This paper will introduce a simple and relative measurement.

3.2.2 Definitions

Definition 1:Coefficient of Variation

The coefficient of variation is the ratio of the standard deviation to the mean.

$$c.v=\frac{\sigma}{\mu} \qquad (2)$$

The advantage of coefficient of variation over other statistical indices is that it is relative and easy to calculate.

Definition 2:Weighted k-clique based on the coefficient of variation (k-clique-w_c.v).

Let G=(V,E,W) be a weighted network, W represents edge weights. A k-nodes complete subgraph $G'=(V',E',W')$ has a total of $\frac{k(k-1)}{2}$ edges. If the coefficient of variation of the $\frac{k(k-1)}{2}$ edges is less than some prescribed threshold $C.V^*$, we call the k nodes a weighted k-clique based on the coefficient of variation. Similarity, weighted k-clique based on deviation is denoted by k-clique-w_σ.

3.2.3 Algorithm Procedure

There are three steps in the process of SR_CPM algorithm:

Firstly, find out all the cliques which are not included in other cliques.

a) Calculate the degree of each node in the network and record the largest value g-1 and then turn to b.

b) Set C to be the collection of all nodes in the network and then turn to c.

c) Construct $C^*=\{v_i|v_i\in C\text{且}d(v_i)\geq m-1\}$ and then take a node v_i from C^* randomly. For v_i, we define two collections A and B. A is a collection of all the nodes which connects each other and contains node $v_i$during the exection of the algorithm, and B is a collection of nodes that are connected to each node in A. Then, turn to d.

d) Iterate recursively to find all the cliques that include v_i and whose size is g. Let $|\mathbf{X}|$ stands for the number of elements in the collection X.

1) Initialize the collection A = $\{v_i\}$, B={the neighbors of v_i}.

2) Each time we move a node from the set B to the set A and adjust the set B by deleting the node that is no longer connected to all the nodes in set A.

3) If $|A|<g$ and $B=\Phi$, or $A\cup B$ is a subset of some clique or $|A|+|B|<g$, stop calculation and return to the previous step in recurrsion.

If $|A|=g$ and $B=\Phi$ and the weight of nodes in set A satisfies definition 2, a new clique is obtained. Record the new clique and return to the previous step in recurrsion and continue to find new cliques.

If $|A|=g$ and $B=\Phi$ and the weight of nodes in set A does not satisfies definition 2, return to the previous step in recurrsion and continue to find new cliques.

If $|A|<g$ and $B\neq\Phi$, execte the (2) recursively.

All the cliques which have the size of g and start from v_i are obtained in the end.

e) Delete v_i from C^*, delete v_i and all the edges connected to it in the network.

f) If $C^*\neq\Phi$, get the next node from the C^* and repeat procedure d ~ e. If $C^*=\Phi$, set g = g-1 ,repeat the procedure b ~ e until g=2.

Secondly, constructing the clique-clique overlap matrix C according to all the weighted cliques found in the previous step.

Thirdly, constructing clique connection matrix and find the k-cliques according to the input k and the matrix C.

3.3 Analysis of the experimental results of the improved algorithm

3.3.1Experimental data

The data set is the classical complex network data set Les Miserables. This data set is a character relationship network constructed according to the relationships among the characters in the Miserable world. The node in the network stands for a character in the novel. If two characters appeared in the same chapter, there will be an edge between two nodes. The weight of the edge stand for the number of times when the two characters appeared in the same chapter. There are 77 nodes and 253 edges in the data set.

3.3.2 Evaluation function

Evaluation function use community partition quality evaluation function—the extension of modular Q which is an evaluation function for overlapping community organizations EQ:

$$EQ = \frac{1}{2M}\sum_{ij}\frac{1}{o_i o_j}\left(W_{ij} - \frac{W_{i.}W_{j.}}{2M}\right)\delta\left(c_i, c_j\right) \quad (3)$$

o_i and o_j represent the number of communities that node i and node j belong to respectively.

3.3.3 Result analysis

The SR_CPM_c.v stands for the SR_CPM algorithm using k-clique-w_c.v and the SR_CPM_σ stands for the SR_CPM algorithm using k-clique-w_σ. For k = 3 and k=4, the results are shown in figure 1.

Under different cv*, the EQ value of the SR_CPM_c.v algorithm for k = 3 is generally larger than the corresponding EQ value for k = 4. This shows the nature that the network is roughly a 3-clique community to some extent. The EQ value increases first and then decreases along with the increase of the variation coefficient threshold cv*, which has a small fluctuation in the middle. The difference of the coefficient of variation cv* shows that the algorithm requires different levels of the degree of dispersion of the edge weights in the clique, that is to say, the degree of familiarity between the users is required. The smaller the cv* is, the smaller the degree of dispersion of

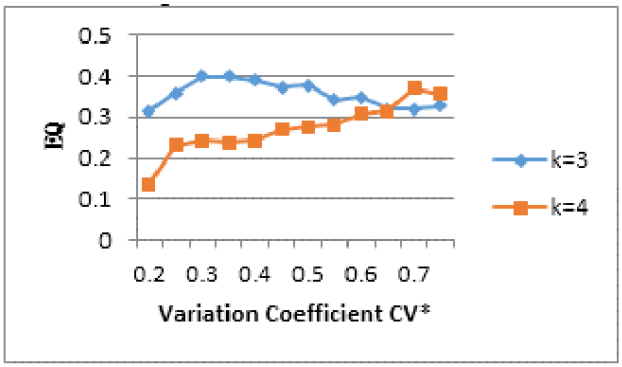

Figure 1. SR_CPM_c.v Change of EQ value under different parameters

the weights will be in the clique and the more severe the conditions for the formation of cliques will be. The larger the value of cv* is, the discrete degree of edge weights requirement in the clique is more broadly. When cv* is too small, some real community structures may not be considered cliques. Although the connections inside the community are very close, the edges among communities is not sparse and then the EQ value is still low. With the

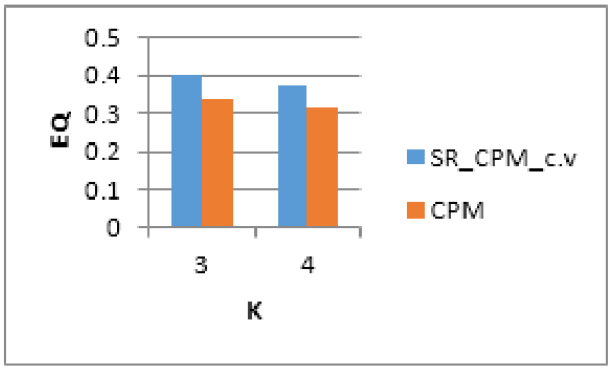

Figure 2.The comparision between SR_CPM_c.v and CPM

increase of the cv* value, the formation condition of the cliques is gradually broad and the EQ value increases gradually as more reasonable nodes join the community. When the cv* is too large, the restrictive condition of constructing a clique in weight is becoming weaker and weaker. Thus, some complete subgraphs with different edge weights are also considered cliques and the EQ value decreases gradually. The algorithm is close to CPM algorithm to some degree without considering the intensity information of the node.

The coefficient of variation of C.V and standard deviation to measure the degree of dispersion are introduced into the definition of cliques. The experimental results show that the SR_CPM is more reasonable in the division of community structure compared to CPM.

4 Application of SR_CPM Algorithm in Telecommunication data Set

4.1 Experimental data

The data set used in the experiment is coming from a telecommunications company data record. There are 3 files including the user information, the call log records and the SMS records. The dataset has been transformed to protect the privacy of the users. The original data contains 382779 users, 76907842 call records and 20947956 SMS records. After preprocessing, the two data set Call_mess_table and User_info are shown below.

Table 1. CALL_MESS_TABLE

Serial number	Attribute name	Type	Meaning
1	user_a_nbr	varchar(8)	phone number of user a
2	user_b_nbr	varchar(8)	phone number of user b
3	all_raw_dur_ab	int	length of the call between a and b
4	call_times	int	times of the call between a and b
5	mess_num_ab	int	short messages between a and b

Table 2. USER_INFO

Serial number	Attribute name	Type	Meaning
1	acc_nbr	varchar(8)	Phone number of the user
2	gender_n	int	Is gender unknown
3	gender_f	int	Is sex female
4	gender_m	int	Is sex male
5	urban_id_0	int	Is it the city
6	urban_id_1	int	Is it the county
7	urban_id_2	int	Is it rural
8	urban_id_3	int	Whether the urban and rural
			identity is unknown
9	certi_latn_551	int	Is the identity card attributable to 551
10	prob_level_100	int	Whether the amount of consumption in the 0 to 100
11	prob_level_200	int	Whether the amount of consumption in the 100 to 200
12	prob_level_300	int	Whether the amount of consumption in the 200 to 300
13	prob_level_800	int	Whether the amount of consumption in the 300 to 800
14	cust_level	numeric(10,2)	User level

The SR_CPM and CPM algorithm are used to partition the community in the two data sets respectively.

4.2 Experimental Content

Two data sets are extracted in this experiment: data set 1 is user call network 1 which has 623 nodes and 3391 edges consisting of a record of 27913 calls from the 623 users and 5624 SMS records. Data set 2 is user call network 2 which has 2403 nodes and 14094 edges consisting of a record of 98152 calls from the 2403 users and 14094 SMS records.

Community discovery is an unsupervised learning process. The structure of the network is not known in advance, so the final results of the community need to be evaluated. There is no authoritative evaluation index of community partition which can be applied to any kind of network at present. The research on the quality evaluation function of the community partition with the weighted network is very little compared with the non-weighted network. It is not easy to evaluate the quality of weighted complex networks. There are a variety of defects in the weighted network when the community evaluation index of many kinds of unauthorized network is applied to the weighted network. Due to the lack of effective social evaluation index of doubt and the weighted complex network classification information, it makes the objective evaluation of the weighted network community discovery algorithm results become extremely difficult. The commonly used evaluation indexes include clustering

coefficient, strong and weak associations and modularity. This paper is still using EQ evaluation function.

4.3 Result Analysis

The SR_CPM algorithm divides the data set into 33 societies (overlapping community structure), and the CPM algorithm divides the data set into 61 societies (with overlapping community structure) with the parameters k = 4 and m = 2.The relationship between community size ranking and community size is shown in figure 3 and 4.

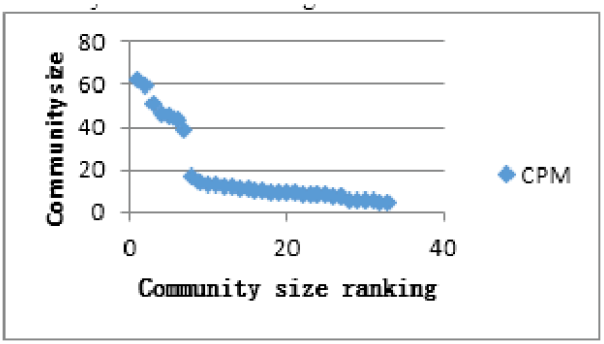

Figure 3 .The distribution of Community size ranking of CPM

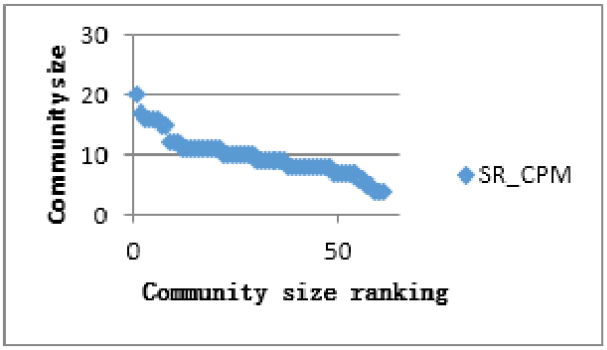

Figure 4.The distribution of Community size ranking of SR_CPM

The SR_CPM algorithm proposed in this paper is better than the CPM algorithm in the case of reasonable parameter settings, and it can find the reasonable community structure in the weighted network. However, if the parameters are set so that the formation conditions are too harsh, the quality of community partition may be inferior to the corresponding parameters of the CPM algorithm. In addition, the division of the SR_CPM algorithm to get the community is different according to the input of different parameters. Telecom operators can adjust them to meet the needs of practical applications.

Acknowledgment

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

References

[1] Girvan M,Newman M E J. Community structure in social and biological networks[J]. ProcNatlAcadSci USA,2002,99(12):7821-7826
[2] Farrahi K,Gatica D. Discovering daily routines from large—scale mobile data[J]. Proceeding of the 16th ACM international conference on Multimedia,2008:849-852.
[3] Pothen A,Simon H,Liou K-P.Partitioning sparse matrices with eigenvectors of graphs[J].SIAMJ Matrix Anal Appl,1990,11(3):430—452.
[4] Newman M E J,Girvan M. Finding and evaluating community structure in networks[J]. Phys Rev E,2004,69(2):026113.
[5] Palla G,Dernyi I,Farkas I,et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature,2005,435(7043):814-818.
[6] Lancichinetti A,Fortunato S,Kertesz J. Detecting the Overlapping and Hierarchical Community Structure in Complex Networks[J].New Journal of Physics,2009,11:033015
[7] Newman ME, Leicht EA. Mixture Models and Exploratory Analysis in Networks[J]. Proc Natl Acad Sci USA,2007,104(23): 9564-9569
[8] Evans T,Lambiotte R. Line Graphs, Link Partitions, and Overlapping Communities[J]. Physical Review E,2009, 80(1):16105.
[9] Farkas I, Palla G, Vicsek T. Weighted network modules[J]. New Journal of Physic, 2007, 9(6):180-198.
[10] Kim H S,Yoon C H. Determinants of subscriber chum and customer loyalty in the Korean mobile telephony market[J].Telecommunications Policy,2004,28(9):751-765