Detecting Social Circle in Social network Automatically

YU Zhanlong[1], DONG Lixin[2], CHEN Yulin[1], FU Chunyan[1,a*], ZHOU Hong[1],ZHI Yuan[1],QU Silong[1]

[1]College of Information Science & Electronoic Technology, JIAMUSI University, Jiamusi 154007, PR China

[2]No.8 Middle Schoole, Jiamusi 154007, PR China

[a]jmsfu@126.com

Keywords: social circle; similarity; overlap

Abstract. Organizations of large and messy personal social networks are challenging issues. However, Construction of such a social circle will take the user a lot of energy, in addition, with the expansion of the user's social circle, the cost of updating timely is amazing. This paper presents a novel machine learning methods to identify a user's social circle automatically. We combined network structure and user profile information to develop a model for detecting social circle. Experimental results proved that the proposed model can accurately identify attribution of diverse data collection in social circles.

Introduction

Online social networks allow users to track the flow of information that hundreds of friends and acquaintances published. User's friends have a lot of information, resulting in the problem of information overload when their organization of their personal social networks. Almost all of the social networks provide such features, such as "list" of Google and the "circle of friends" of Weixin and Renren. Circle of friends that user-created can be used for content filtering (such as filtering the status information that acquaintances updated),privacy protection (such as hiding personal information to specific person), tracking information of user groups that is concerned.

At present, users of Weixin, Google, and Renren either classify their social circle manually, either determine friend by common attributes. Both methods are less satisfactory. This paper studies how to find the user's social circle automatically, especially given a user's personal social network, how to determine his social circle. Every circle of friends is a subset of his friends, social circle is user-specific, because the friends of each user's social circle are independent of the user's who are not associated with him. This means that you can describe the social circle detection as the relationship clustering problem between his personal network and the network and of his friends'. Shown in Figure 1, a specified individual user u, his friend vi form a network, node vi is defined as a variable point. Task of this paper is to determine which set that vi is belongs to, and then find the nesting and overlapping clusters in personal network.

Network in Figure 1 shows a typical behavior which can be observed from the data directly: about 25% of the aggregation (obtained from Weixin) is completely contained in another gathering circles, 50% of the aggregation overlap another aggregation, 25% of the aggregation have no intersection with other gathering circles. Goal of this paper is to find those aggregation through the network relationships between personal friends, to discover aggregation members and find common attributes of this gathering circle.

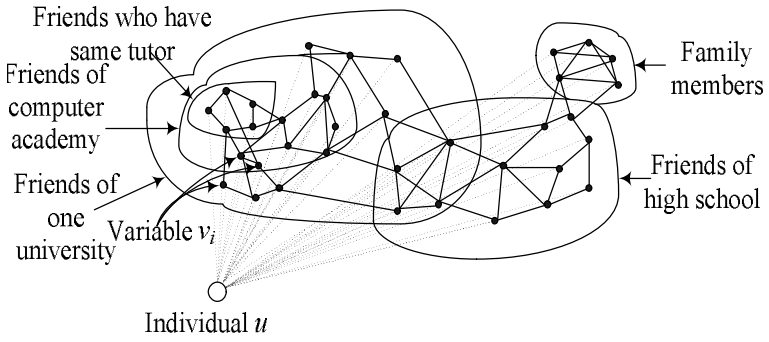

Fig. 1 define the gathered personal network

This article contributions are as follows: 1) The propose of prediction method that node is difficult to assigne to crowded circle, solve the difficult problem that models such as Mixed Member models can not handle[1], to obtain good performance; 2) put forward similar arguments personal information definition and form links in accordance with similar dimensions, allows the different social dimensions constitute different aggregation, expands the concept of homogeneity.

The generation Model of friends relations in social network

Model of circle of friends should follow the following properties:

(1) nodes in the cluster should have common attributes or characteristics;

(2) different clusters should constitute different characteristics, an aggregation may be composed by family members,another aggregation may be composed by students of an university;

(3) clusters should be allowed to overlap, and a "strong" set can form in the "weak" set, for example, the circle composed by people who have same degree can be included in the circle of friends in the same university, shown in Figure 1.

(4) we should take adervantage of personal information and the structure of the network to determine the cluster. it should be exactly knowed that which aspect of the information that constitutes a cluster, such the model is illustrated to the user. This paper presents a model to describe the friend relationship in the social networks according to above analysis.

The input of the model is a personal network G=(V ,E) ,and information of each user v(v∈ V), The center of the personal network u is not included in G and G only contains the friends of u (variable points). The reason to define a personal network in this way is because the creator of the circle of friends are not in these circles. Each aggregation set in private network is C={C1…Ck} , Ck $\subseteq$V , the relevant parameters vector indicates how aggregation occurs, the information of user is encoded into a tuple features $f(x,y)$, capture the co-properties of the x and y in a certain way.

The circle model consideres members of the circle as potential variables.Nodes Fall in public circles usually have the opportunity to form a side, it will naturally lead to layering and overlapping of social circles. The paper integrate the latent variables and similar parameters of information, design unsupervised algorithms in order to explain the observed network data better. Given a personal network G ,and K social circles sets C={C1…Ck}, defines a set of points (x, y) $\in V\times V$ and the probability of an edge may be formed is:

$$p((x,y)\in E)\propto \exp\{\underbrace{\sum_{C_k\supseteq\{x,y\}} < f(x,y),q_k >}_{\text{circles containing both nodes}} - \underbrace{\sum_{C_k\cup\{x,y\}} a_k < f(x,y),q_k >}_{\text{all other circles}}\} \tag{1}$$

For each aggregation,we will get the similar parameters of information. If two nodes both belongs to C_k ,then the value of $< f(x,y) , q_k >$ will be very big, and the value will be very small if any point does not belongs to C_k (a_k balances these two effects). As the information similarity between the user x and y is defined by the feature vector, the parameters vector defines which information similarity of the dimension forms the aggregation circle. are generated independently, Considering edges $e=(x,y)$ the probability of G can be defined as:

$$P_\Theta(G;C)=\prod_{e\in E} p(e\in E)\times\prod_{e\notin E} p(e\notin E) \tag{2}$$

$\Theta=\{(q_k,a_k)\}^{k=1...K}$ is the model parameter. Define shorthand notation:

$$d_k(e)=d(e\in C_k)-a_k d(e\notin C_k)$$

$$\Phi(e)=\sum_{C_k\in C} d_k(e) < f(e),q_k >$$

So the log-likelihood values of G can be written as:

$$l_\Theta(G;C)=\sum_{e\in E}\Phi(e)-\sum_{e\in V\times V}\log(1+e^{\Phi(e)}) \tag{3}$$

Then,we will descripte how to optimize the members C of the circle, and the user information $\Theta=\{(q_k,a_k)\}(k=1...K)$。

similar functions $\Theta=\{(q_k,a_k)\}(k=1...K)$ when graph G and the user information are given.

Experiments

Experimental data sets

In order to evaluate the unsupervised algorithms on real data sets, the paper access to personal networks and real data from three major social networks: Weixin, Google and Renren, including 193 friends circle and 4039 users.

We have developed a special Wechat application program which was used to investigate the 10 users, requiring them to manually determine which circle of friends of their friends should belong to.On average, users will probably determine 19 circles of friends, containing 19 members averagely. These circles of friends can be classmate friends, sports teams, relatives and so on.

The two kinds of the eigenvectors about the information are 1- $s_{x,y}$ 和 1- $s'_{x,y}$:

1- $s_{x,y}$ =[0 0 1 1 0 1 1 0 0]^T , The meaning of each corresponding item to the vector is as below: [Name: Li xiaoming, Name: Wang fuxing, Position of work: Password designer, Company: The

Guoxin Systems Company, Company: ZTE Company, the institution of the education on Education situation: Southeast University, Education type on Education situation: undergraduate, the institution of the education on Education situation: Hehai University, Education type on Education situation: graduate student].

1- $s'_{x,y}$ =[0 1 1 1 1]^T ,The meaning of each corresponding item to the vector is as below: [Name, Position of work, Company, Education type on Education situation , the institution of the education on Education situation, Education type on Education].

We get 133 individual network from the Google + network, involving 479 circles of friends and 106674 users. These 133 personal networking represents all share at least two circles of friends in the Google +from 133 users, what's more, the information of these people's network is public.

At last, from the 1000 Renren personal networks, including 4896 circles (or column [19,27,31,10]) and 81362 users, we choose the 10 to 4964 nodes according to the size of the personal network. All the data are composed of 1143 individual network, 5541 social circle and 192075 users.

The data form Wechat is completely marked among them, and users in essence believe that community circles of friends are cohesive, while the data from the Google + and Renren net is marked partially, that is, public circle is only used.

Experimental content

Although the proposed model is a non-supervised learning, but the maximum likelihood value $C=\{C_1...C_K\}$ can be detected by real data convergence potential circle of friends, and the goal is that the potential circle of friends and the manual suggests circle should be closed in great extent for a proper regularization model. In order to measure the proximity of C and $\bar{C}$,we calculate the balance error rate (BER) [2] of two sets of bit, $BER(C,\bar{C})=\frac{1}{2}(\frac{|C\setminus\bar{C}|}{|C|}+\frac{|\bar{C}\setminus C|}{|C|})$. In this approach, a pseudo-real and a pseudo- false is equally important, so that the average error caused by subtle or random forcast can be about 0.5.

Because we don't know the alignment of C and $\bar{C}$, we get the optimal matching by calculating linear maximum value

$$\max_{f:C\to\bar{C}}\frac{1}{|f|}\sum_{C\in dom(f)}(1-BER(C,f(C))) \tag{4}$$

Here f is corresponding to C and $\bar{C}$, thas is, if the number of C($|C|$) is less than the number of $\bar{C}$ ($|\bar{C}|$), then there will be a matching $\bar{c}\in\bar{C}$ for each $c\in C$, but if $|C|>|\bar{C}|$, there is no additional matching. Also we can estimate the number of friends by using the maximum likelihood estimate technology.

In this paper, we compare this method with the following three methods, the first one is the multitasking clustering algorithm put forward by Streich[3], we say that clustering; the second is Low-Rank algorithm put forward by Yoshida[4], we note for the Low-Rank; the third is block-LDA algorithm put forward by Balasubramanya and Cohen[5], we note for the LDA.

In this paper, we note for "F2F12" when this method is running under the condition of friends

on friends ($\square^1$=12), “F2U13” in the case of friends on users ($\square^2$=13) , and note for “C114”when this method is running under the condition of compression characteristic (ψ^2=14) .

Figure 2 shows comparison results when this algorithm tests the accuracy of community in micro letter , Google+, renren data sets.

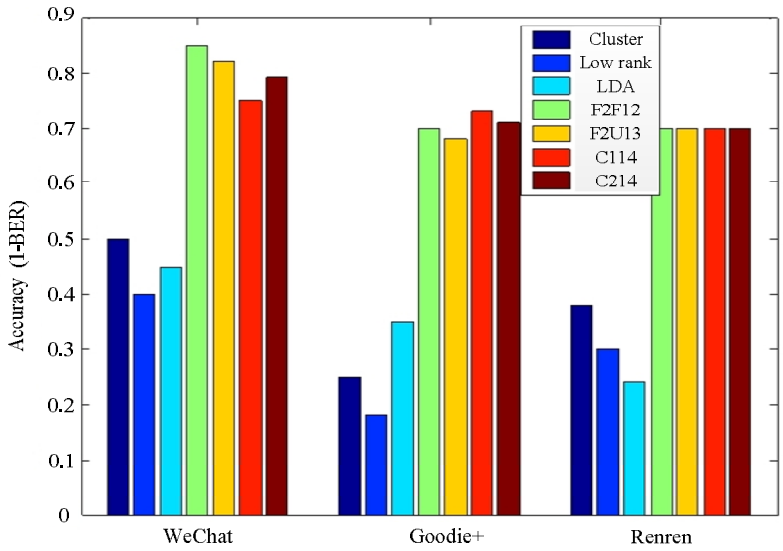

Fig.2 Compare the accuracy of each algorithm detecting communities

We describe the circle of friends by the type (4), we get the number of circle according to the type (3). If the balance error rate (BER) is higher, the performance will be better. The article columnars denote standard error rate in figure three.

The difference between the running accuracy of this method under the condition of best eigenvalue □ case and the nearest competitor is 1%. At this point the scores of BER is: micro message is 0.84, Google+ is 0.72 and renren is 0.70.

The reason why Google+ and renren get low score is: Since the initial users are created, there does not hold many social circles, and can achieve high recall value (to regain friends in each circle), but the prediction precision is low (the additional friends after the circle of friends established.

We can see that the good performance of this method mainly relies on the combining sites and edge information to predict members of the multiple friends circles from the experimental results. Up to now there is no other method applying on this combination.

Conclusion

The paper proposes a method that performs aggregate operation on social data. The method can learn in completely unsupervised condition and can determine the aggregation number and their members.The paper collected 1143 personal web data sets from the micro letter, Google and renren, getting 5636 manual true classification of different social circles. The experimental result of the network data sets shows that the method of considering the social network structure and users' personal information at the same time is obviously better than the natural selection and the currently popular method. The method also can explain why the nodes belong to some gathering at the same time of improving the accuracy.

Acknowledgements

This paper is funded by the science and technology research project of Hei Longjiang Education Department (12531678)。

References

[1] E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic blockmodels[C]. JMLR, 2008.
[2] Y. Chen and C. Lin. Combining SVMs with various feature selection strategies[C]. Springer, 2006.
[3] A. Streich, M. Frank, D. Basin, and J. Buhmann. Multi-assignment clustering for boolean data[C]. JMLR, 2012.
[4] T. Yoshida. Toward finding hidden communities based on user profiles[C]. In ICDM Workshops, 2010.
[5] R. Balasubramanyan andW. Cohen. Block-LDA: Jointly modeling entity-annotated text and entity-entity links[C]. In SDM, 2011.