Online Community Perceiving Method on Social Network

Jingchi Jiang[1], Chengqi Yi[1], Yuanyuan Bao[2], Yibo Xue[2*]
[1]School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, 150080, China.
[2]Tsinghua National Lab for Information Science and Technology, Tsinghua University, Beijing, 100084, China.
(jiangjingchi@mail.tsinghua.edu.cn, garnettyige@163.com, by51800@163.com, yiboxue@tsinghua.edu.cn)

***Abstract*—How to perceivecommunity with the same interests and hobbies on social network is a critical problem for advertisement, promotion and network security. In this paper, we propose a novel method for online communityperceiving, which contains three user-level filtering layers: user profile, user behavior and user relationships.This three-layer filtering mechanism can greatly improve the accuracy and efficiency of onlineperceiving specific community. In order to testify this method, we design a web crawler forfurther determining the identity of specific nodesby breadth-firststrategy.The experiments onTwitter Chinese users demonstrate that the precision of our online community perceiving method can reach up to 89%.**

Keywords- online community perceiving; social network; web crawler;filtering mechanism

I. INTRODUCTION

With the rapid development of the Internet in recent years,network community is increasingly popular with users,such as Facebook, Travel Blog and Twitter. According to the figures released by an Analysis Organizationcalled Semiocast,Facebook has over 1.2 billion users and generates more than 70 billion pieces of contents per month. Twitter has over 0.5 billion users and generates 190 billion tweets every day [1].

With the increasing amounts of information and users on social network, it is impossible to analyze the entire dataset. Given the characteristics of community on social network, obtaining concerning data from a specific community is a viable option.Meanwhile, considering the direct impacts on the value of information, the timeliness of obtained information is another influencing factor. So how to perceive specific community effectively and promptly is crucial for commercial promotion, public opinion monitoringand terrorism detection.

In this paper, the main contributions are as follows:

- We propose a novel filtering mechanism for online community perceiving, which contains three filtering layers: user profile, user behavior and user relationships;
- In order to testify the three-layer filtering mechanism, we design a web crawler according to the mechanism to further determine the identity of specific community, which is obtained by breadth-firststrategy;
- We evaluate our method by doing experiments in perceiving Twitter Chinese users by API, which demonstrate the precision and efficiency of our method.

The rest of this paper is arranged as follows.InSection IIwe discuss the previous research on community detection. In Section III, we propose an online community perceiving method based on a three-layer filtering mechanism. And also define the principle of each layer. In Section IV,we conduct experiments to testify the efficiency of proposed perceiving method.In Section V, we conclude this paper and discuss the directions for further work.

II. RELATED WORK

Community detection algorithm can be roughly divided into two categories: non-overlapping and overlapping. Non-overlapping detection algorithm refers to identify the non-overlapping community, in which each node only belongs to a unique community. The non-overlapping algorithms include modularity optimization algorithm [2-4] and spectral analysis algorithm [5]. In modularity optimization algorithm, community detection is defined as an optimization problem, and the aim is changed into searching the targeted optimal community structure; the main idea of spectrum analysis algorithm is to determine the number of community and identify the community structure by eigengaps and eigenvectors of adjacency matrix.

Aforementioned non-overlapping community detection algorithmsdivide each node strictly to a single community, whichis not consistent with the actual situation. In reality, people often belong to multiple communities.Therefore, the overlapping community detection becomes a new hotspot in the area of community detection in recent years. One algorithm of overlapping community detection is based on fuzzy clustering, in which the strength of community's membership to each community can vary and can be expressed as a belonging coefficient that describes how a given vertex is distributed between communities.And this kind of fuzzy analysis algorithm usually starts from calculating distances between nodes [6].

Traditional community detection methods play important role in community detection. However, most of methods are implemented on stable offline datasets without timely updates, which will significantly affect the accuracy and timeliness of the results. In this paper, we adopt an initiative online community perceiving method, which can perceive a specific community on social network with timely online data. The method can greatly improve accuracy and

[*]Corresponding author. Tel: +86-010-62772393. E-mail: yiboxue@tsinghua.edu.cn

efficiency of specific community perceiving, having great significance for dynamical community detection.

III. Online Community Perceiving Method

As an important medium of communication and information dissemination, social network is made up of a set of social users with different interests and a complex set of bilateral ties between these users.Social users communicate with others having the similar background, behavior and interests. Due to this characteristic of convergence, communities with different features are formed. The emergence of community is an inevitable phenomenon and user-level characteristics, such as user profile, user behavior and user relationship, have significant impacts on community formation.Based on these analyses, we design a three-layer user-level filtering mechanism to ensure the accuracy of community perceiving. As is shown in Fig.1, perceiving method is roughly divided into three layers: a) The first layer is designed to match the string pattern of the user profile with keywords extracted manually from specific community; b) Considering the behavioral features, the second layer is designed to calculate the similarity of behavior features between the detected user and members of the community; c) By considering the feature of users relationship, the third layer is designed to compare clustering coefficient of users with those of community. When the former is greater than the latter, the detecteduser will be joined into the specific community.

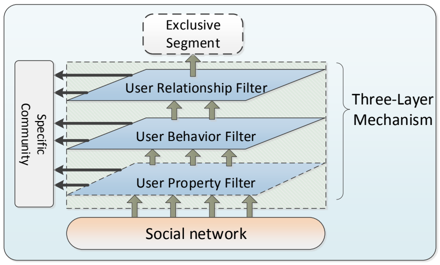

Figure 1. The architecture of perceiving method

A. User ProfileFeatures Filter

On social network,user profile includes properties such as name,email, introduction,etc. Most of these are filled by users themselves.So personal profile can reflect some features of users, which can provide strong characteristics for community perceiving. Artificially abstractinga list of keywords, which called as “description wordsof community”,is an important step.Then,the first layer filter adopts a classical algorithmon fast string matchingcalled KMP algorithm to match the user profile with the list of keywords. Ifuser’s introduction or username contains keyword, it illustrates the relevance between user and perception community. Then the filter puts user into specific community.

The user profile layeradoptingstring matching methoddepends on the characteristics of user profile and keywordsof the community. This layercan quickly and accurately achieve the filter of the basic information.

B. User Behavior Features Filter

However,the featureof user profile is not enough to complete the whole filtering task. In this situation, another filter based on user behavior feature is proposed.

User behavior refers that the information is produced by subjective will of social users,which can accurately reflect the user's attention.We use Vector Space Model (VSM) to calculate the similarity between users and community.

In order to calculate the similarity, we establish an n-dimensional vector based on community description words as shown in Eq.(1), where$T$represents the frequency of a keyword in community,Nrepresents the number of selected community description words.

$$U = [T_1, T_2, T_3, \ldots, T_N] \quad (1)$$

Then,weusesentence segmentation to deal with all the texts of user, which satisfy the condition in the first layer.The value of keyisthe frequencyof eachdescriptive word in user’s text messages. We can calculatethe similarity between user and specificcommunityby Eq. (2)-(3).

$$P_A = \left[key_1, key_2, key_3, key_4, ..., key_N\right] \quad (2)$$

$$sim_{(A,U)} = \cos_{(P_A,U)} = \frac{P_A \square U}{\|P_A\|\|U\|} \quad (3)$$

By the value of$sim_{(A,U)}$,we can measure the relevance of node and community.When the value is greater than a threshold, the model putthe node intothecommunity. Threshold is crucial and should be determined through large amounts of experiments. After the node joinsto community,the description word shouldbe updatedalong with the changes of keywords list of community.The dynamical process can dig up some potential description words,which can make up for the disadvantages that description wordsof community arecompletelyartificial.

C. User Relationship Feature

The third layer usesthe local clustering coefficient,acharacteristic of strong community, to perceivewhethera node belongs to the specific community.

Strong community: If communityζsatisfies the following condition, $\zeta$is called as strong community [7].

$$k_i^{in}(\zeta) > k_i^{out}(\zeta), \forall i \in \zeta \quad (4)$$

$k_i^{in}(\zeta)$represents the connections of node i within the communityζ,while $k_i^{out}(\zeta)$ represents the connections of node i with the rest of communityζ.

Clustering coefficient:The clustering coefficient can be divided into two categories: node coefficient and community coefficient [8].

a) The node coefficient is defined as the proportion of connections among its neighbors which are actually realized compared with the number of all possible connections. The parameter k is defined as the number of connectionsbetween node i and communityζ.

$T(i)$representsthe number of all possible connections among the k vertices.

$$T(i)=k(k-1)/2 \tag{5}$$

$E(i)$represents the actual number of edges among the kvertices.$c(i)$is the clustering coefficient of node i and can be computed as follows.

$$c(i)=E(i)/T(i) \tag{6}$$

b) The community coefficient is defined as the mean of the entire node coefficient within the community. $c(i)$ is defined as the clustering coefficient of community ζ:

$$c(\zeta)=\frac{\sum_{i-1}^{n}c(i)}{n} \tag{7}$$

If a node conforms to the characteristics of strong community, we conclude that the node belongs to the specific community. However, if the condition is not satisfied, we further calculate the clustering coefficient of adjacent nodes.The bigger the node coefficient, the higher the connectivity between the node and specific community.According to the above analysis, we can affirm whether node belongs to the specific community.

Based on the above descriptions of three filtering layers, we can achieve an online community perceiving method with three-layer user-level filtering mechanismshown in the following table.

Algorithm 1 Online community perceiving

```
Input$i$: the initial target node
Output$\zeta$ : specific community
Begin
Initialize the feature vector to perceivecommunity
    $P\{f_1, f_2, ..., f_n\}$
    Select the initial node $i$ , then put it into$\zeta$
while $j \in \zeta$ & $j \neq null$
    Using crawler to get node set with the link to $j$
    on internet，define as List$_j$
for each $v_j \in List_j$
if the attribute of includes a certain feature of
$P\{f_1, f_2, ..., f_n\}$ then $v_j \to \zeta$
else if $sim(v_j, U(\zeta)) >$ thresholdthen
$v_j \to \zeta$ andupdate$U(\zeta)$
else if $k_i^{in}(\zeta) > k_i^{out}(\zeta)$ || $C(i) > C(\zeta)$
                then $v_j \to \zeta$ and updatec$(\zeta)$
end if
end for
end while
```

Following the analysis above, online community perceiving method can be implemented. Throughtestifying the adjacent nodes of initial target nodewhich belongs to the specific community adopting the three-layer filter, we can determine whether the testified node belongs to the specific community. If the node satisfies the condition of any layer, we put it as part of the community. According to this filtering mechanismcontinuously, the specific communitycan be obtained.

IV. EXPERIMENTS AND EVALUTION

In order to verify the effectiveness and efficiency of online community perceiving method, we conduct experiments to perceive a community consisted of Chinese people on Twitter. The statistics and analysis of experiments mainly rely on three indicators including cosine similarity threshold selection, validity of method and the time consumption of method.

A. The Sensibility ofCosine Similarity Threshold

In the experiments of this section, we measure the threshold of cosine similarity inbehavior features filter.

When the cosine similarity which iscalculated based on user's tweets and keywords list is greater than a fixed value, the targeted user is put into the Chinese community.We have done multiple sets of experiments to determine threshold, the experimental result is shown in Fig.2. The accuracy reaches highest when the threshold is 0.008, after that the accuracy declines slowly. So we choose the threshold as 0.008 in our method and system.

In addition, the quantity of tweets is another factor to determine the accuracy. Due to the Twitter API limitation, crawler can obtain a maximum of 200 tweets at each time. So we compare the accuracy of 20 tweets and 200 tweetsas shown in Fig.2.From Fig.2, we can see that the accuracy of 200 tweets is much higher than that of20 tweets. By considering the Twitter API limitation, therefore we choose 200 tweets for computing.

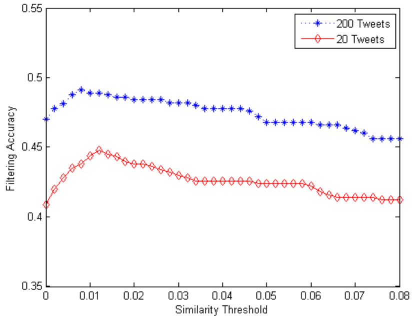

Figure 2. Cosine similarity thresholdselection chart

B. *Method Accuracy Analysis*

In order to testify the accuracy of this method, we choose a dataset with 40,000 users including 2,000 Chinese. We firstly extract dataset into three-layer filter. The dataset is composed of 2,000 Chinese and other country users in proportion of one to one. Because the quantity of Chinese remains unchanging, the accuracy will be influenced bythe increasingamounts of other users. Considering the actual situation, we add 2,000 users to the former dataset in each time. As the proportion of Chinese is adjusted fromone to nineteen, the accuracy always remains stable and unchanging as shown in Fig.3, no matter how the proportion changes.

In addition, the experimentsare done to testify the accuracy of method in different levels. When perceiving method just relies on the user profile layer, the accuracyis 20%, as shown in Fig.3.If user behavior layer is added, the accuracy of method will be improved, which can reach up to 60%. Furthermore, if user relationship layer is also added, the method improves the precision obviously, which is close to 90%. In conclusion, our extensive empirical studies on the real datasets show the effectiveness of our approach and the necessity of multi-layer filter.

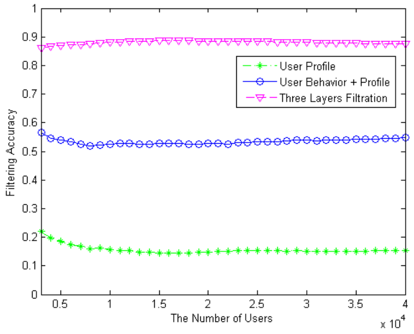

Figure 3. The accuracycomparison of using different level

C. *Time Consumption of Method*

In order to verify the efficiency,we need to further analyze the time consumption of this method.Three histograms respectively representtime consumption in each layer of filter.

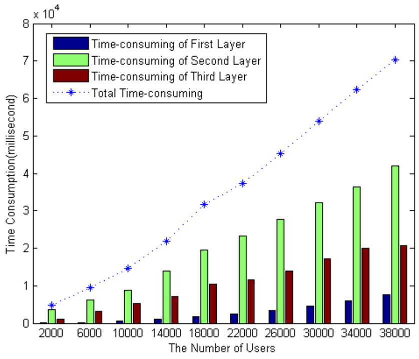

Figure 4. The time comparison of using different level

From Fig 4,the user behavior filter has the highesttime consumption, in whichtext segmentation and similarity comparison are time-consuming. Through statistics of timeconsumption of each layer,the total timewhich is represented by the broken line increases linearly with the increase of the quantity of dataset. The time for filtering 38000 users is about 70 seconds, it can reach real time perceiving, thus demonstrates effectiveness of our approach.

V. CONCLUSIONS

In this paper, weare researching ononline community perceiving methodon social network. Based on three-layer user-level filtering mechanism including user profile, user behavior and user relationships as the characteristics of each layer, we propose a novel method for online community perceiving.This three-layer filtering mechanism can greatly improve the accuracy and efficiency of online specific community perceiving. In order to testify this method, we design a web crawler to further determine the identity of specific communityby breadth-firststrategy.The experiments of TwitterChinese users demonstrate the precision of our online community perceiving method can reach to 89%. Our future work will focus on optimizing the algorithmof filter and cutting down time consumption in the process of perceiving.

ACKNOWLEDGMENTS

This work was supported by the National Key Technology R&D Program of China underGrant No.2012BAH46B04.

REFERENCES

[1] AOL Inc. Analyst: Twitter passed 500M users in June 2012, 140M of them in US; Jakarta ‘Biggest Tweeting’ city[EB/OL]. (2012-07-30) [2012-10]. http://techcrunch.com/2012/07/30/analyst-twiter-passed-500m-users-in-jun-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/.

[2] Zhou Z, Wang W, Wang L. “Community Detection Based on an Improved Modularity,”Pattern Recognition. Springer Berlin Heidelberg, 2012: 638-645.

[3] Lancichinetti A, Fortunato S. “Limits of modularity maximization in community detection,” Physical Review E, 2011, 84(6): 066122.

[4] Zhang X S, Wang R S, Wang Y, et al. “Modularity optimization in community detection of complex networks,” EPL (Europhysics Letters), 2009, 87(3): 38002.

[5] Ruan X M, Sun Y H, Wang B, et al. “The Community Detection of Complex Networks Based on Markov Matrix Spectrum Optimization,” Control Engineering and Communication Technology (ICCECT), 2012 International Conference on. IEEE, 2012: 608-611.

[6] Sun P G, Gao L, Shan Han S. “Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks,” Information Sciences, 2011, 181(6): 1060-1071.

[7] Radicchi F, Castellano C, Cecconi F, et al. “Defining and identifying communities in networks,” Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(9): 2658-2663.

[8] Bródka P, Filipowski T, Kazienko P. “An introduction to community detection in multi-layered social network,”Information Systems, E-learning, and Knowledge Management Research. Springer Berlin Heidelberg, 2013: 185-190.