Behavior-Based Twitter Overlapping Community Detection

Lixiang Guo(✉), Zhaoyun Ding, and Hui Wang

College of Information Systems and Management,
National University of Defense Technology, Changsha, 410073, China
{guolixiang10,zyding,huiwang}@nudt.edu.cn

Abstract. In this paper, we try to cluster twitter users into different communities. These communities can be overlapping based on their interests. The paper proposed a RWC (relation-weight-clustering) model to construct twitter users' network. This model takes twitter users' "@" and "RT@" behaviors into account. By counting their "@" and "RT@" frequency, the relation strength can be then descripted. Using SVM, we can get the users interest vector by analyzing their tweets. And the common interest vector between two users is calculated according to their common interests. Using community detection algorithm to resolve the relation-nodes-based network, the overlapping communities are formed with modularity of 0.682.

Keywords: Overlapping community detection · Interest space · RWC model

1 Introduction

Many more social web services like Twitter, Facebook, LinkedIn, etc. which are based on social network have emerged in the latest decade. It has attracted many researchers to investigate the mass data generated by them every day.

Community detection has been a hotspot in the field of social network research. In other words, the community detection technology enables us to find community structure in the social network and to get a deep insight of relations or interests among nodes.

After Michelle Girvan and Mark Newman proposed the concept of modularity [1] in 2002, community detection really took off. Many algorithms have been put forward aimed to optimizing the modularity function. A typical one of them is FN algorithm, a greedy optimization method. It views every nodes as small independent communities at the beginning. And then combine two communities as a new one, where these two communities are really linked in the network and the value of the modularity increases most or deceases least. After the end of iteration, all the nodes become a community. The modularity is calculated in each iteration. Then the community partition which is corresponding to the largest modularity value is the approximately optimal community structure. Besides, the modularity is often used to evaluate the quality of community structure generated from other community detection algorithms.

H. Gao et al. (Eds.): DASFAA 2016 Workshops, LNCS 9645, pp. 371–376, 2016.
DOI: 10.1007/978-3-319-32055-7_31

In fact, users in the Twitter-liked social network may have kinds of interests. This means a user can belong to more than one interest-based community. Hence the community structure should be overlapping. To detect overlapping community, Ahn *et al*. [2] proposed LCA algorithm which reinvented communities as groups of links rather than nodes. The groups of links were mapped to nodes at last. Then the overlapping community was gotten. Zhou *et al*. built an R-C model [3] taking the link similarity into consideration to improving LCA.

The work of Zhou *et al*. does not consider the relation strength between users. This might lead to unreasonable community structure. On the foundation of their work, we take the relation strength into account. And the community detected is more reasonable.

In this paper, we collect tweets from 47360 Twitter users. User interest space is built through analyzing their tweets' contents. By counting the frequency of their "@" and "RT@" behaviors, the relation strength can be easily got. Finally, the modularity of the overlapping community is 0.682.

2 Relation-Weight-Clustering Model

The relation link between two users is represented as a common interest vector. Then the weight is added to the common interest vector. We view the weighted common interest vector as clustering object. Using fast optimizing algorithm [4], the relation links are clustered into several groups. Finally, the relation links are mapped to user nodes, which is corresponding to user communities.

2.1 User Interest Vector Construction

A user may usually have different interests. It means that a user may belong to different interest-oriented communities. We use support vector machine (SVM) to gain the users' interests. A user's interest vector **I** is an n-dimension vector, where $I = (w_1, w_2, \ldots, w_n)$ and n is the number of interests. Each dimension of **I** is a specific interest and its value is the possibility of the user's tweets on this interest.

2.2 Relation Link Interest Vector Construction

We assume that two users become friends for sharing common interests. Based on this assumption, we use a vector **C** to represent the relation link between two users. And **C** is defined as

$$C = I_i \cap I_j, \tag{1}$$

where

$$I_i \bigcap I_j = (\min\{w_{i1}, w_{j1}\}, \min\{w_{i2}, w_{j2}\}, \ldots, \min\{w_{in}, w_{jn}\}). \tag{2}$$

Considering that the interaction frequency reflects the relation strength intuitively, we think about adding an interaction-related factor ω to **C**. The definition of ω is

$$\omega = \max\{\alpha_1 \frac{\#U_1@U_2}{\#U_1@} + \alpha_2 \frac{\#U_1RT@U_2}{\#U_1RT@},\ \alpha_1 \frac{\#U_2@U_1}{\#U_2@} + \alpha_2 \frac{\#U_2RT@U_1}{\#U_2RT@}\}, \tag{3}$$

where $\#U_1@U_2$ is the times of user $\mathbf{U_1}$ "@" user $\mathbf{U_2}$ in all tweets of $\mathbf{U_1}$ and vice versa. $\#U_1@$ is the times of all the "@" behavior of $\mathbf{U_1}$'s tweets and vice versa. And "RT@" denotes retweet behavior. α_1 and α_2 are the weight which are satisfied

$$\alpha_1, \alpha_2 > 0, \alpha_1 + \alpha_2 = 1. \tag{4}$$

Therefore the weighted relation link interest vector can be

$$C_w = \mu \cdot \omega \cdot C, \tag{5}$$

where μ is an alterable factor for adjusting the scale of C_w to an appropriate level.

2.3 The Relation-Based Network Construction

We use R-C network model [3] as the basic model. In this network, the nodes are relation links as above. There is an edge between two relation links if and only if they share common user. The weight of edge can be regarded as the similarity of those two relation links. Hence it can be defined as

$$W(C_{w1}, C_{w2}) = \frac{C_{w1} \cdot C_{w2}}{\left|C_{w1}\right|^2 + \left|C_{w2}\right|^2 - C_{w1} \cdot C_{w2}}. \tag{6}$$

The equation above is *Tanimoto coefficient* (also called *Extended Jaccard coefficient*). It is easy to find that $W(C_{w1}, C_{w2})$ is between zero and one. A larger $W(C_{w1}, C_{w2})$ means that the two relation links are more similar.

3 Experiment and Analysis

3.1 Data Collection

We have collect 47360 Chinese twitter users' profiles and their tweets in September, 2015 by using twitter API. And the data are stored in the MySQL database.

3.2 Data Processing

Twitter User Interest Vector. For a twitter user, we analyze every tweets of him. Those tweets with too many non-Chinese characters are filtered out.

At first, we decide to classify these tweets into six categories (*A*, *B*, *C*, *D*, *E* and *Other*). Actually, tweets which are in *Other* group are also neglected. Using SVM, the rest of tweets are attached with a unique label.

Then, a user's interest vector can be represent as

$$I = (\frac{\#A}{total}, \frac{\#B}{total}, \frac{\#C}{total}, \frac{\#D}{total}, \frac{\#E}{total}), \tag{7}$$

where $total = \#A + \#B + \#C + \#D + \#E$.

Twitter Users Relation Link Interest Vector Construction. We extract "@" and "RT@" relations from tweets. If two users have "@" or "RT@" behavior each other, there will be a link between them. And the relation link interest vector can be calculated by using (5).

Table 1 shows a pair of users with its weighted interest vector. The adjusting factor is set to 1000.

Table 1. Some examples for relation link interest vector

u1(id)	u2(id)	Weight	Relation interest vector	Weighted vector
127262132	2236766378	0.136	0.17, 0.08, 0.05, 0.2, 0.22	23.12, 10.88, 6.8, 27.2, 29.92
1265070655	1862357449	0.196	0.14, 0.05, 0.13, 0.18, 0.18	27.44, 9.8, 25.48, 35.280003, 35.280003
1265070655	2833539408	0.043	0.14, 0.02, 0.44, 0.08, 0.15	6.02, 0.85999995, 18.92, 3.4399998, 6.4500003
870309318	1618790083	0.034	0.15, 0.04, 0.4, 0.12, 0.14	5.1000004, 1.36, 13.6, 4.08, 4.76
870309318	16865364	0.027	0.15, 0.04, 0.19, 0.12, 0.16	4.05, 1.0799999, 5.13, 3.24, 4.3199997
870309318	145440266	0.048	0.15, 0.03, 0.53, 0.11, 0.11	7.2000003, 1.4399999, 25.439999, 5.2799997, 5.2799997
870309318	633328589	0.014	0.15, 0.04, 0.41, 0.12, 0.16	2.1000001, 0.56, 5.74, 1.68, 2.24
18190842	2197807908	0.014	0.24, 0.06, 0.11, 0.2, 0.19	3.36, 0.84, 1.54, 2.8, 2.6599998

3.3 Twitter User Relation-Link Network Construction

There is an edge between two relation links if and only if they share common user. Then the weighted undirected network is constructed as in Table 2 below.

Table 2. Some examples for twitter user relation-link network structure

LinkId_1	LinkId_2	tanimotoSim[a]
3	4	0.293
4	14006	0.996
8	15856	0.3
8	15865	0.296
9	54000	0.486

[a]"tanimotoSim" is the tanimoto coefficient.

3.4 Community Detection

Clustering the Relation-Links into Groups. Using maximizing modularity method in [4], the twitter user relation-links are partitioned into 471 communities with modularity of 0.682.

Mapping the Relation-Links to Twitter Users. *Rule 1*. If some relation-links are in the same group, users attached to these relation-links belong to a group.

Based on rule 1, the final twitter users overlapping communities are detected as in Table 3. The numbers in the community column are the IDs of communities. That a user corresponds to several IDs means the user belongs to these communities at the same time.

Table 3. Some samples for overlapping communities of twitter users

userId	Community
1001077530	153
100122533	454, 115, 232, 169, 104, 25, 236, 88, 360
1001268486	152, 115, 360
100172757	171
100175420	117, 48, 128, 115, 55, 169, 33, 168, 277, 32, 40, 358, 104, 88, 141
100176531	104
100198190	171, 470, 148, 241, 136, 55, 195, 400, 169, 168, 360, 141
100233785	262, 300, 115
100253361	115, 188, 88
100506067	354, 183, 40

4 Conclusion

This paper does not take friendship relation among twitter users as the source of basic network. We take the "@" and "RT@" behaviors among twitter users as the basic composition of network instead. It is reasonable to do that because the online interactive behavior ("@" and "RT@") can reveal the common interests even they are strange each other in the real world. Using RWC model, we find the overlapping communities based on interest.

This method can be applied to different scene to get high quality and reasonable communities in the social network.

References

1. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E: Stat. Nonlin. Soft Matter Phys. **69**(2Pt2), 026113–026113 (2004)
2. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010). doi:10.1038/nature09182
3. Zhou, X.P., Liang, X.: User community detection on micro-blog using R-C model. J. Softw. **25**(12), 2808–2823 (2014)
4. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. **30**(2), 155–168 (2008)