Mining Users for Organizations on Micro-blogs

Zhenhua Zhang[1] **Ruifang Liu**[1] **WeiRan Xu**[1]

[1]Beijing Univversity of Posts and Telecommunications

Abstract

This paper focuses on detecting relevant real-world users of organizations in micro-blogs. Many organizations and individuals choose to publish posts on micro-blogs. As the virtual social network may possibly be influenced by accounts' real world social ties, it's feasible to detect the user groups that have strong connections to organizations. In this paper, we propose a community-detection based method to discover the real world members of organizations on micro-blogs. In order to find the most relevant user groups, we also define ranking score for each user. Experiments show that, the top ranked users of each community is crucial for distinguishing the most relevant groups.

Keywords: Micro-blog, Community Detection, User Analysis, Organization.

1. Introduction

Micro-blog is one of the most successful innovations based on Web 2.0 in recent years, popular micro-blog platforms like Twitter, has reached more than 500 million registered users in 2012[1].On the network, we can easily distinguished some organization-based accounts apart from the individual-based ones. According to Java et al[2], there are generally three types of micro-blog users: information source, friends or colleagues, and information collector. Most of the organization-based accounts are among the first category, whereas individual users create accounts base on various purposes.

This paper proposes a novel approach based on community detection to mine the relevant members for organization-based accounts, that is, discover the micro-blog user groups who might have strong correlation to organizations in the real world. For instance, employees may give rise to relevant communities to their company, students can form circles that related to their schools and so on. As far as we know, there's no previous research about real member mining on micro-blog.

The community or circle mentioned here refers to group that are more densely connected internally than with the rest of the network[3]. There are generally two kinds of behaviors that could contribute to the relevance of a user to organizations, i.e. either posting relevant tweets or following relevant accounts to the organizations. For most users, posting relevant tweets is only temporary behavior, users will lose their relevance after they quit from the discussion. Moreover, discovering all relevant tweets to organizations on twitter is another sophisticated problem which we won't discuss here. This paper will mainly introduce an approach based on users' social relationship. As users' follower and following information is far less time-dependent, the relevant user groups mined based on user network is relatively more stable.

The paper is build up as follows: section two describes the collection of the user relationship data for our experiments; section three introduces the ranking strategy of relevant users to the organizations; section four presents the methods that we used to find the relevant communities and evaluate experiment result; Finally, we make concluding remarks in section five.

2. Data Collection

On micro-blogs, users usually choose to follow other users that they are interested in. From a user’s following and follower information, we can easily discover a number of most relevant users to the given account.

In this paper, we exploit the above property of the micro-blog social network to find the relevant users. On the network, three types of users are necessary for the mining of relevant real world members:

- official accounts: a set of users created by the given organization or its subordinate groups.
- first level followers: a set of users who follow at least one official user and are not among the official accounts.
- second level followers: a set of users who follow at least one first level follower and not are among the official accounts or first level followers.

The official accounts serve as the representatives of the organization on micro-blog, while the first level followers could be the most relevant user group as they follow the official accounts by interest, some real world members of the organizations are expected to be found among them. In fact, not all the real world relevant users of the organization will choose to follow the official accounts, but they may be friends with their colleagues or schoolmates on micro-blogs, so the second level followers also possibly include some relevant members. To mine the relevant users, we build the user network by crawling the following and follower information of all official accounts and the first level followers, then extract all bi-directional edges, for bi-directional edges usually means the two users are friends in the real world or at least they have strong interest in each other. We crawl the user relation data via the Twitter REST API[4]. To accelerate the crawling and reduce the size of the data set, we eliminate users with more than 10,000 followers, those users are usually celebrities or news agencies and could easily introduce a large number of irrelevant second level followers.

For the tested organization, it has 49 official accounts, 11,663 first level followers and 4,609,126 second level followers. The directed graph consists of 8,213,899 vertices and 23,967,287 edges. As for the undirected graph, it consists of 1,697,230 edges and 941,961 vertices, that means there are more than nine hundreds of users connected from the official accounts by bi-following relationship. As micro-blog users may participate in numerous social circles, the graph unavoidably contains a great number of noisy nodes.

3. Ranking the Relevant Users

As the following information usually reveals user’s interest on social network, users that follows a big number of relevant users to the organizations usually has strong interest in the organization. For every user in the network, we measure its relevance base on its following distance to the official accounts. Users’ relevance score is defined as:

$$RScore_i = \frac{1}{l_i \sum_{j=1}^{|U_0|} \frac{1}{d_{ij}}} \qquad (1)$$

l_i is the distance that user u_i reaches any official account, d_{ij} is the length of the shortest path from user u_i to official account u_j. U_o is the set of official accounts. User's relevance score is dependent upon relevance of the users they follow, Users who follows more official accounts or users of high RScore are sufficiently relevant to the organizations, those users are more unlikely to follow the official accounts at random on the user graph. The calculation of the RScore of all first level and second level is extremely fast, and the experiment shows that the distribution of user's RScore follows a power law distribution.

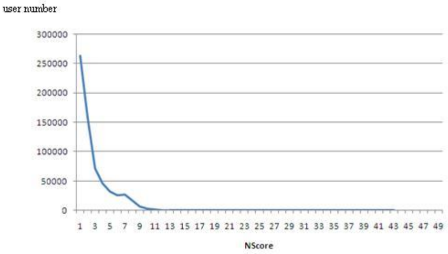

Fig. 1: user RScore distribution

Though the RScore is reliable in ranking users based on their connectivity to the official accounts, it's not enough to discover the real world members. Because the real world relevant users usually are expected to belong to a related social community, whose members have bi-directional relationship with other users and may display collective relevance to the organizations.

4. Relevant Communities Detection

To analysis the network, we use the fast greedy algorithm proposed by Clauset, Newman and Moore[5], it has good balance on speed and accuracy, and is particularly effective in handling massive data set. However, the bi-directional network still extremely enormous to implement the algorithm, so we continue to cut off the network by eliminating both first level followers that follow only one official accounts, and second level followers that follow less than two first level accounts. Those users are usually noises of the graph and slow the community detection procedure. This method is simple but extremely effective in reducing the size of the graph. The reduced graph has only 34,660 vertices and 56,685 edges. The graph yields 18 communities whose sizes is between 500 and 5,000, along with a number of small and noisy groups.

But the above communities are not all relevant to organizations, some may only are connected to the official accounts via the bi-directional edges, We still have to select the most relevant communities after the community detection step. To achieve this goal, We define the relevance weight of the communities as:

$$Weight_i(k) = \frac{N_{ki}}{\sum_{j=1}^{|C|} N_{kj}} \qquad (2)$$

N_{kj} denotes the number of users that belong to community c_j among the top k users with the highest RScore. With a proper k(usually between 500 and 2000), the relevance weight can effectively indicate the relevance of the group. We demonstrate the validity of the method by checking the user information of the communities.

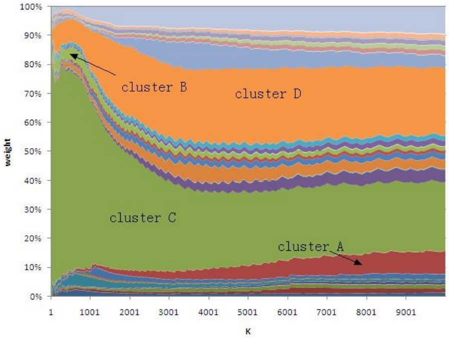

Fig. 2: communities' relevance weight

We select four sample communities among all user groups to illustrate the result of the experiment. From the beginning of K axis of Figure 2, cluster C and cluster D are two major relevant groups according to their relevance weight. In fact, there are 36 official accounts in cluster C, nearly all members of the community share the same geo location with the test organization. After ranking all users in this group on their RScore, we could easily figure out that lots of the top ranked users are real world members of the test organization from their personal description. Cluster D includes 2 official accounts, also includes some real world members, though not as many as cluster C. In cluster B, only a small number of top ranked users are relevant enough. As for cluster A, it keep expanding as k increasing, which suggests that it consists of mainly low RScore users. Obviously, cluster C and cluster D are quite relevant communities, while cluster A is not. Figure 2 show the RScore of the four representative communities. As Cluster C includes 36 official accounts(official accounts' RScore is |Uo|, i.e. 49), its top ranked members exhibit obvious advantage over the other sample communities.

The relevance weight emphasizes the relevance of the most relevant users of each community, it proofs that the top ranked users' relevance is highly effective in finding the most relevant communities. It can be noted from Figure 2 that the RScores of all four communities decrease sharply with the growing of K and quickly turn into a stable value afterward. In fact, in each community, there are still a large number of users that don't display any relevance from either their user information or RScore. Those users could be friends or relatives of the relevant users, but are not real world users of the organization. This means that the graph still keeps a lot of noise. A further de-noising approach is to prone nodes with low RScore in each community with certain threshold, only keep the most relevant users.

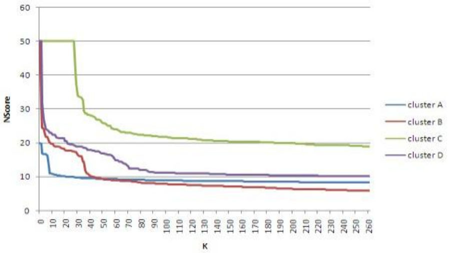

Fig. 3 top ranked users in the sample communities

5. Conclusion

In this paper, we propose a new approach based on community detection to mine real world relevant members on microblog platform. All bi-directional relationship are kept to find all true friends on the network. We also remove the noisy users according to their following information. After detect the user communities, we define communities' relevance weight to distinguish the most relevance user groups. Observation shows that the relevant groups still includes lots of noise, which could be eliminated according to their RScore.

6. References

[1] http://en.wikipedia.org/wiki/Twitter.

[2] Java, A., Song, X., Finn, T., & Tseng, B. Why we Twitter: Understanding microblogging usage and communities. WebKDD/SNA-KDD '07 Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM New York, NY, USA. 2007 .Pages 56-65

[3] http://en.wikipedia.org/wiki/Community_structure

[4] https://dev.twitter.com/docs/api

[5] M. E. J. Newman (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E* **69** (6): 066133. 2004