Mining Social Relationships in Micro-blogging Systems

Qin Gao, Qu Qu, and Xuhui Zhang

Department of Industrial Engineering, Tsinghua University. 100084, Beijing, China
gaoqin@tsinghua.edu.cn

Abstract. The widespread popularity and vigorous growth of micro-blogging systems provides a fertile source for analyzing social networks and phenomenon. Currently, few data mining tools can deal with unique characteristics of micro-blogging systems. In this study, we propose an integrate approach for mining user relationships in micro-blogging systems. The approach starts from macroscopic analysis of social networks by grouping users with the method of maximal strongly connected components (MSCC). Following that, a measure of condensation level of groups are calculated to find out the most influential group , and all groups can be ranked according to this measure; then a new algorithm is presented to evaluate the influence of a specific user within a group. The integrated approach is capable to analyze large amount data sets. It is useful for exploring directions of information diffusion and evaluating the scope and the strength of individual user's influence in micro-blogging systems.

Keywords: Social data mining, micro-blogging systems, information diffusion analysis, and graph mining.

1 Introduction

Micro-blogging systems, like twitter.com, let users write a brief text about daily life and update immediately through many different ways, including typing online, text messaging, instant messaging (IM), and sending emails. In recent years, micro-blogging systems become increasingly popular, and the user scale is huge. A report found that in June 2010, nearly 93 million internet users visited Twitter.com, an increase of 109 % from the previous year[1]. The large volume of user's social connection information captured by such systems provides many opportunities for mining social networks and phenomenon. In particular, in a micro-blogging system, users need to explicitly indicate whether they want to hear from another user by "following" or not. Knowing such explicit and directed relationships will allow not only structural analysis of social networks but also enables the study of information flow within the networks. Furthermore, networks in micro-blogging systems overlap heavily with social networks in real life. Many users are familiar with their followers

[1] http://www.comscore.com/ger/Press_Events/Press_Releases/2010/8/Indonesia_Brazil_and_Venezuela_Lead_Global_Surge_in_Twitter_Usage

A.A. Ozok and P. Zaphiris (Eds.): Online Communities, HCII 2011, LNCS 6778, pp. 110–119, 2011.

and followees not only in virtual world but also in real life [1], and new users generally join an existing network by accepting friends' invitations. This makes social relationship analysis in micro-blogging systems more indicative of real social networks than analysis in other online communities.

So far, most research of user relationships analysis studied blog communities, social network sites (SNS), and collaborative tagging systems [e.g., 2-4], and few studies micro-blogging systems. Existing methods used in these studies, such as Social Network Analysis (SNA), have limitations in analyzing huge volume of data sets, and most of them cannot incorporate information flow directions into the analysis. In this paper, we propose an integrate approach to mine social relationships in micro-blogging systems. In addition to high efficiency that is required to process huge data size, our approach also provides a tool for exploring the directions of information dissemination between users and evaluating individual users' influence in information dissemination.

The rest of this paper is organized as follows: The following section describes related studies and motivation. Section 3 introduces the new integrated approach. A case of implementation of our method is presented in Section 4. Section 5 shows our conclusion of the work.

2 Related Work

2.1 Analysis of Online Social Networks

The popularity of online social network services makes social relationships mining come again into the limelight with the new communication platform. A lot of researches explore web user relationships in blog, SNS or other web communities from different perspectives with topological analysis [6], link analysis [7] and network evolution [8].

SNA is one of the most influential methods in social relationships analysis. Aligning centrality measures are adopted to discover communities in blogs [2], and betweeness measures are used to extract natural community structures of social networks by dividing the network nodes into densely connected subgroups [3]. Tyler et al. [9] proposed an algorithm based on betweenness and centrality to discover user groups in email networks. Previous studies show that most of SNA researches emphasized binary interaction data, with direct and/or weighted edges, and they focused almost exclusively on very small networks [10]. With the development of internet, especially the rapidly spread of micro-blogging systems, the limitations of SNA become more and more obvious. The size of data set is booming. SNA, however, is hard to process a huge data set. Besides, the traditional SNA is difficult to explore the directions of information transmission in micro-blogging systems.

There are a few of other methods of online social relationships mining. To mine a directed social network from an online message board, Matsumura et al. [11] simplify the algorithms of Influence Diffusion Model (IDM) [12] in which the influence of a user is evaluated by propagating terms among couples via messages. Kazienko and Musiał [13] present a new method of personal importance analysis to discover personal social features in the community of email users based on calculation of the strength of relationships between network members, its dynamic as well as personal position of the nearest neighbor. Similarly with SNA, these methods have difficulties in processing large volume of data and analyzing directed information flow.

2.2 Graph Theory

Graphs are widely employed as general data structures in modeling complex systems and networks. Mining frequent subgraph patterns is an effective way to research characters of graphs. There are two basic approaches for pattern analysis: the apriori-based approach and the pattern-growth approach. Apriori-based approach begins with a small size frequent subgraph, and proceeds in a button-up manner by generating candidates with big "size" frequent subgraphs having an extra edge, vertex, or path [14]. The apriori-based approach is based on breadth first search, and they require a good-sized system working space. Different from the the apriori-based approach, the pattern-growth approach adopts depth first search and this approach expand 'small size' subgraphs by adding new vertexes or edges.

Relational graph is a special graphic structure in which each vertex is unique. Relational graphs are frequently used for modeling biological network, social network, traffic analysis and internet analysis. Dense subgraph is a type of relational graphs, which usually denotes close relationship within a group. CluseCut and Splat are two main algorithms of frequent dense subgraphs mining [14].

Some new methods of analyzing huge directed networks are brought from graph theory. Samudrala [15] used graph theory to discover protein structures. The algorithm denotes each possible conformation of a residue in an amino acid sequence with the notion of a node in a graph. Each node gets a weight based on the degree of interaction between its side-chain atom and local main-chain atoms, and draw edges between pairs of residue conformations/nodes that are consistent with each other [15]. Cai et al. [16] proposed a regression-based graph matrix approach to explore hidden communities in heterogeneous social networks, and they validated the approach with the Iris and Digital Bibliography and Library Project (DBLP) datasets.

3 An Integrate Approach of Social Relationships Mining

3.1 A General Information Diffusion Model

Garton et al. [17] suggested a regular social network can be described as a finite set of nodes that are linked with one or more edges, and we build an information diffusion model to explore the user network and information flow in micro-blogging systems. In the model, nodes indicate users or user groups, and directed edges represent information diffusion between two users. Based on this model and existing research related to mining of huge and directed graph data, a new integrate approach based on this model is introduced in the following.

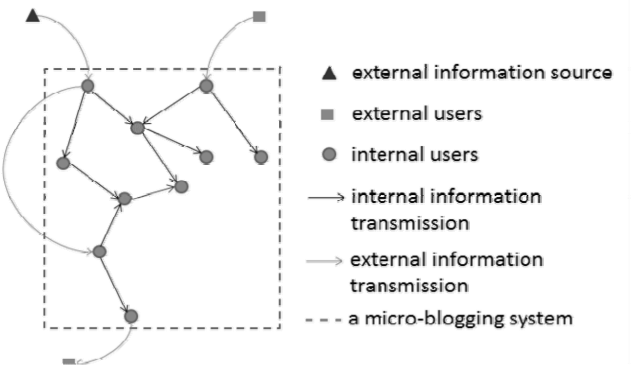
▲ external information source
■ external users
● internal users
→ internal information transmission
→ external information transmission
- - - a micro-blogging system

Fig. 1. Information diffusion model

3.2 Step 1: User Grouping by Information Dissemination Relationships

Micro-blogging users spontaneously form a large number of groups by various interests or different regions. For social relationships mining, an ideal grouping is that information exchange within a group is frequently while information exchange with members out of a group is as less as possible. Based on information diffusion model, we adopt maximal strongly connected components (MSCC) to group users in micro-blogging systems. A MSCC in micro-blogging systems can be defined in this way: Given a directed graph G = (V, E), where V is a finite set of nodes, and E (E $\in$ V×V) is a finite set of directed edges. Nodes denote users, and directed edges express information flow. For $\forall a \in \mathrm{V}$, $\forall b \in \mathrm{V}$, if there is at least one path between *a* and *b*, then the directed graph G is a strongly connected component. And if G would not a strongly connected component when any node or edge were added to G, it is a maximal strongly connected component (MSCC).

Based on the definition of MSCC, we defined a user group as a set of nodes within which any two nodes can transfer information bidirectionally. That is, any two users can’t transfer information bidirectionally except for that they belong to the same MSCC.

3.3 Step 2: Group Ranking by Information Dissemination Paths between Groups

Obviously, the importance of groups in information dissemination is different. For this reason, we propose a method to rank groups according to their contributions in information dissemination.

Each group is denoted as a node, that is, the internal relationships within a group are masked and only the relationships among groups are visible. The network of a micro-blogging system is then condensed into a directed acyclic graph G´ (Details of proof are showed in Appendix), and each node of G´is a MSCC. We adopted a topological sorting algorithm to rank groups in G´with necessary modifications. The node without any information outflow is deleted from G´and put at the end of the ranking list. This step is repeated till all nodes are deleted. In the final ranking list, the first one is the strongest influential group in the micro-blogging system. The pseudocode of this algorithm is showed as follows:

```
P<Set<Node>> ←Empty list that will contain sets of
nodes in sequence
N ←Set of nodes with no outside link

Insert all nodes which have no outside link into N
while N is non-empty do
   insert N into P
   for each node n in N
     remove n
       for each node m with a link e from n to m do
           remove e
```

3.4 Step 3: User Influence Estimation by the Probability of Information Dissemination

In addition to ranking the influence of groups, knowing the strength of a specific user has on his/her followers within a group, especially in the most influential group, would be interesting from a practical point of view. Dijkstra [5] proposed a well-known algorithm to explore the single-source shortest path problem for a directed graph, and the algorithm is often used for finding costs of the shortest paths from a node to another node. Inspired by Dijkstra's algorithm, we introduce the concept of width and propose a new index using the name of QIndex to estimate the influence of a certain user in a group.

In micro-blogging systems, information transmits via subscriptions (by following others) or retweeting by others. We define the number of nodes from the source node to the target along a path as the distance of that path and define the number of different paths connecting the two nodes as the width. The probability that a piece of information reach the target depends on the length and the width of information transmission paths. The shorter the distance and the more paths existing between two nodes, the more probably information can reach the target node. We assume P as the probability that any user retweets a certain update. Therefore, the probability that the target user can receive this information is $p= \Sigma_{i\in N} P^{d_i}$, where N represents the set of all paths from the source to the target, and d_i is the length of path i. According to observation, it is reasonable to infer that P < 0.5. Therefore, the shortest path from the information source to the target node makes the greatest contribution to p. To simplify the problem, we can set a threshold T and if d_i is larger than T, p_i which reflects the probability that information transmits via path i can be considered approximating to zero. Thus we only consider the shortest path within T.

Based on the above inference, we propose an algorithm based on Dijkstra's to calculate the influence of a specific node. For a directed graph G = (V, E), V is a finite set of nodes and E is a finite set of edges and E $\in$ V×V. The information source node is labeled as v_s ($v_s \in$ V). The algorithm is described in following steps:

1. Set the distance value to zero for initial node v_s and to infinity for all other nodes, and assign a width value to one for v_s and to zero for all other nodes. Mark all nodes unvisited, and set v_s as the current node, noted as n_c.
2. Then we need to calculate the distance and width between the current node n_c and all unvisited nodes to which it links. Assume the distance from the source node to the current node is d_c, and the width of the current node is $w_{c,}$ n' is an unvisited node which is linked by *the current node.* Before this round of calculation, the distance and width between the source node and n' is d' and w', and the distance between n' and the source node via n_c is d_c +1. If $d_c +1 < d'$ and $d_c +1 <$ T, then the distance between the source node and n' will be replaced with d_c +1 and w' will be updated with w_c at the same time; if $d_c +1= d'$ and $d_c +1 <$ T, w' will be changed into $w'+1$.
3. The current node n_c will be marked as a visited node when all of its unvisited nodes are calculated.
4. Set the shortest distance node in all unvisited node as current node n_c, and repeat step 2. If there isn't any unvisited nodes in a distance less than T, QIndex of all visited nodes will be calculated and the algorithm will be finished. If width=0, the QIndex is infinity, otherwise, the formula is

$$\text{QIndex=Distance/Width} \quad (1)$$

The higher the QIndex, the less probably the target node would receive information from the source node. It is very important to set a proper threshold T based on required accuracy, computing resources and other limitations. In the worst case, the running time of QIndex algorithm is O $(|V|^2 + |E|)$. Threshold T is proportional to time cost and the level of detail about the final result, that is, a larger threshold T costs more time and gets more detailed results, while the smaller T costs less time and gets more roughly results. If T approximates zero, the time cost of QIndex is close to O $(|V| + |E|)$.

4 Validation

User and usage data from www.digu.com, a popular micro-blogging system in China were collected. Digu.com was established in February, 2009, and the function and user interface design is very similar with twitter.com.

We use snowball sampling method for data collection. Snowball sampling which is a nonprobability sampling method with which future subjects are recruited from acquaintances of exiting subjects, just like a rolling snowball. Twenty users were chosen randomly from the public discussion board of digu.com as 'seeds', and the information of user and user relationships were collected for each seed (Table 1). Then all the followers and followees of each seed were chosen as 'seeds' again.

Our data collection system sends a request to digu.com every 10 seconds, and the whole data collection lasts two weeks (March 23, 2010 - April 7, 2010). Finally we collected 332,122 users and 11,160,822 inter-user connections. Because of computing resource constraints, a smaller sample includes 2,556 users with 35,510 inter-connections were used in further computation. There are 60 users' relationships changed during data collection. However, the lost data only account for 2.34% of all users so that the effect is limited on the final result.

Table 1. A data collection example

Item	
ID	11528569
User name	ququjoy
Nick name	Qu
Location	Beijing
Gender	1 (1-male, 2-female, 3-private)
Self-introduction	From Chongqing
Address	http://pic.minicloud.com.cn/file/default/SIGN_24x24.png
Homepage	http://digu.com/ququjoy
Information Privacy	false(false-information disclosure, true-information
The Number of Followees	2
The Number of followers	2
The Number of updates	7
Folloee	digu, robot
Follower	xabcdefg, flyinglin456

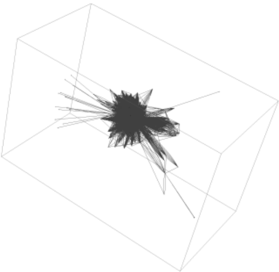

Fig. 1. MSCC of the biggest group

Through calculating MSCC, the biggest group which contains 1,426 users is found.

Then we randomly choose a user named yoohee1221_to find out the top 5 users most influenced by him. We set T=5, and the result of QIndex calculation shows yoohee1221_ has strong influence on his direct followers, e.g., classyuan, which is not a surprising result. However, although some users don't follow yoohee1221_ directly, e.g., liuxinwu, QIndex shows that they are as strongly influenced by yoohee1221_ as those directly linked to yoohee1221_ (Table 2.). Result shows using MSCC and QIndex to mine user relationships of micro-blogging systems is effective.

Table 2. Result of Qindex

Users	Distance	Width	QIndex
classyuan	1	1	1
gambol	1	1	1
liuxinwu	2	2	1
xujun99663	3	2	1.5
dan123	4	2	2
chervun	4	2	2
tuniu	4	2	2
harliger	4	2	2
zxb888	4	2	2
topidea	4	2	2
yuanjuan	4	2	2
WDM123	4	2	2
shaun	4	2	2

5 Conclusion

In this paper, we propose a new approach to mine user relationships in micro-blogging systems. First, the new method of user grouping by maximal strongly connected components (MSCC) is introduced for social network structure analysis. Second, by condensing the graph and sorting with a topological algorithm, groups are ranked according to their influence on other groups. Third, a new algorithm that inspired by Dijkstra's algorithm is presented to assess the influence of individual users on others. The integrate approach can be applied to explore directions of information diffusion in micro-blogging systems and discover opinion leaders, and the results can serve various purposes, including product advertisement, policy advocacy, viral marketing and other information diffusion applications in micro-blogging systems.

References

1. Java, A., Song, X., Finin, T., Tseng, B.: Why We Twitter: Understanding Microblogging Usage and Communities. In: The Joint 9th WEBKDD and 1st SNA-KDD Workshop 2007, pp. 56–65. ACM, New York (2007)
2. Chin, A., Chignell, M.: A Social Hypertext Model for Finding Community in Blogs. In: 17th Conference on Hypertext and Hypermedia, pp. 11–22. ACM, New York (2006)
3. Newman, M.E.J., Girvan, M.: Finding and Evaluating Community Structure in Networks. J. Phys. Rev. 69(2), 26113 (2004)
4. Girvan, M.: Community Structure in Social and Biological Networks. PNAS 99(12), 7821–7826 (2002)
5. Dijkstra, E.W.: A Note on Two Problems in Connexion with Graphs. J. Num. Math. 1(1), 269–271 (1959)

6. Ahn, Y., Han, S., Kwak, H., Moon, S., Jeong, H.: Analysis of topological Characteristics of Huge Online Social Networking Services. In: 16th International Conference on World Wide Web, pp. 835–844. ACM, New York (2007)
7. Hsu, W.H., Lancaster, J., Paradesi, M.S.R., Weninger, T.: Structural Link Analysis from User Profiles and Friends Networks: A Feature Construction Approach. In: ICWSM 2007, pp. 75–80. ACM, New York (2007)
8. Golder, S., Wilkinson, D., Huberman, B.: Rhythms of Social Interaction: Messaging within a Massive Online Network. J. Com. and Tech. 2007, 41–66 (2007)
9. Tyler, J., Wilkinson, D., Huberman, B.: E-Mail as Spectroscopy: Automated Discovery of Community Structure within Organizations. J. Info. Soc. 21(2), 43–53 (2005)
10. McCallum, A., Wang, X., Corrada-Emmanuel, A.: Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email. J. Arti. Inte. Res. 30, 249–272 (2007)
11. Matsumura, N., Goldberg, D., Llorà, X.: Mining Directed Social Network from Message Board. In: 14th International Conference on World Wide Web, pp. 1092–1093. ACM, New York (2005)
12. Matsumura, N.: Topic Diffusion in a Community. In: Ohsawa, Y., McBurney, P. (eds.) Chance Discovery, pp. 84–97. Springer, Heidelberg (2003)
13. Kazienko, P., Musiał, K.: Mining Personal Social Features in the Community of Email Users. In: Geffert, V., Karhumäki, J., Bertoni, A., Preneel, B., Návrat, P., Bieliková, M. (eds.) SOFSEM 2008. LNCS, vol. 4910, pp. 708–719. Springer, Heidelberg (2008)
14. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
15. Samudrala, R., Moult, J.: A Graph-theoretic Algorithm for Comparative Modeling of Protein Structure. J. Mol. Biol. 279(1), 287–302 (1998)
16. Cai, D., Shao, Z., He, X.F., Yan, X.F., Han, J.W.: Mining Hidden Community in Heterogeneous Social Networks. In: The 3rd International Workshop on Link Discovery, pp. 1–26. ACM, New York (2005)
17. Garton, L., Haythorntwaite, C., Wellman, B.: Studying Online Social Networks. JCMC 3(1), http://www.ascusc.org/jcmc/vol3/issue1/garton.html

Appendix

Lemma 1. *Given $G = (V, E)$ and $G' = (V', E')$, where G' is the condensation of G. G' is a directed acyclic graph.*

Proof. Suppose that there is a cycle $\tilde{c} = (\tilde{v}, \breve{e})$ in G' and $\forall \tilde{p}, \tilde{q} \in \tilde{v}$. Then we get $\tilde{p}$ and $\tilde{q}$ are reachable from each other. Thus $\tilde{c}$ is a strong connected subgraph of G', which is contradictory to the condition that G' is the condensation of G. Therefore, G' is a directed acyclic graph.

Lemma 2. *Given a directed acyclic graph $G = (V, E)$. If $\exists\, n \in V$, there will not be any outside link from node n.*

Proof. Suppose the size of set V is L and n_q has an outside link to n_q. If$n_q$has not any outside link, the lemma will be proved; if n_q has an outside link to nr, it is easy to prove that *r≠p≠q* for the reason that G is a directed acyclic graph. Continue this iteration until n_r has no outside link, then the lemma will be proved. After L times iterations, n_L will be get. We hypothesize there is an outside link from n_L to n_i. Because G is a directed acyclic graph, the inference that $i \neq 1$, …, L and the hypothesis that the size of V is L become contradictory. Therefore, there is at least one node in G which has no outside link.