Visualization Research of the Tweet Diffusion in the Microblog Network

Jing Lu   Xiaoqing Yu   Wanggen Wan
School of Communication and Information Engineering
Shanghai University
Shanghai, China
yxq@staff.shu.edu.cn

Jing Lu
School of Electronics and Information Engineering
Shanghai University of Electric Power
Shanghai, China

***A*bstract—With the development of big data technology, information visualization plays an increasingly important role in the social network, especially the microblog network. The path of the tweet diffusion shows the relationships among microblog users and constructs a directed network among them. The microblog data mining project and the visualization process of the tweet propagation are described in detail. The visualization results of the tweet propagation indicate that there are two typical transmission path patterns: the dandelion pattern and the double-star pattern. The visualization research may easily discover the key retweeting nodes in the tweet diffusion network.**

Keywords—microblog; visualization; retweeting; tweet diffusion;path

I. Introduction

Microblog service has emerged as a new medium. Users can post tweets within a limit of 140 characters to share information, discuss ideas and exchange opinions. One of the popular microblog platforms is Twitter, which is commonly adopted almost all over the world. In China, Weibo, a Chinese-version microblog service, has played an important role in people's life. According to the official statistics from China Internet Information Center (CNNIC), up to the end of December 2013, the number of microblog registered users is 281 million. More and more people like the "grassroots media"---Weibo.

In the microblog network, the follower-friend relationship is the essence of the information diffusion and the retweet function is the key element. If user B follows user A in the microblog, B is called a follower of A, while A is called a friend of B. User B can read all the tweets created or retweeted by A without any approvals[1]. If user B sees a tweet that he feels would be interesting or helpful to his followers, he can just retweet the post with/without comments by simply pressing the retweeting button. The message will diffuse extensively after several rounds of retweeting. The path of the tweet transmission shows the relationships among microblog users and constructs a directed network among them[2].

The information in the microblog network is massive, complex and unstructured, so traditional data analysis methods have been difficult to adapt to these features. The visualization analysis of microblog data using visualization tools is a very powerful research direction and has a broad prospect of application. It can make the boring microblog data become vivid and complex microblog data relationships become clear.

Researches have built some visualizations for analyzing social network and microblog data from different domains. Jose[3] presented network visualization as s way to analyze and to test the design of a website. Patrick[4] proposed a platform which could provide a SNA(social network analysis) visualization for dark network analysis. J.Ratkiewicz[5] researched political discourses on twitter from three perspectives: networking topics, networking media objects and networking actors and showed the visualization diagram of the interaction network of Twitter users discussing the Pelinka case. Hakan Kardes[6] obtained a better picture of organization collaborative patterns in the funded research network. These scholars put the visualization research emphasis on the static member structure of the network, while the visualization of the information propagation is dynamic and more complex in structure. Ref.[7] built a system which consisted of two interfaces: a web-based online visualization interface for public users and an offline expert visual analytic system and could visualize Weibo events, but unfortunately, it didn't show the specific process of data mining.

In this paper, the visualization method of the tweet diffusion in the microblog network will be developed in the aim of better understanding the microblog network. The paper is organized as follows. In Section II, we will simply introduce the concept of the information visualization and the Gephi platform. The data mining scheme in the microblog network will be described in Section III. Section IV will show two propagation path modes in the microblog network and their visualization results. Finally, we will discuss the visualization research results and what can be further expected in Section V.

II. INFORMATION VISUALIZATION AND GEPHI PLATFORM

A. Information Visualization

Data analysis is an indispensable part of all applied research and problem solving in practical activities. The most fundamental data analysis approaches are statistics, visualization, data mining and machine learning methods. Among these approaches, information visualization is the most reliant on the cognitive skills of human analysts, and allows the discovery of unstructured actionable insights that are limited only by human imagination and creativity.

Information visualization is the study of visual representations of abstract, nonphysical data to reinforce human cognition. The abstract, nonphysical data includes both numerical and non-numerical data, such as text, hierarchies and geographic information. It is not only about creating graphical displays of complex and latent information structures. It also contributes to a broader range of cognitive, social and collaborative activities.

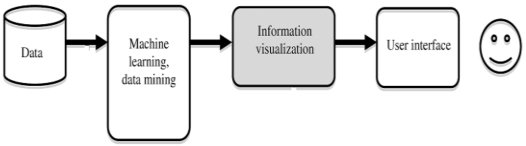


Fig. 1. Conceptual illustration of information visualization

Fig.1 shows the conceptual illustration of information visualization. Information visualization is based on the data. To obtain useful knowledge from “big data”, information visualization must aid machine learning and data mining methods which have made considerable progress in recent years. Meanwhile, the user interface for the successful application of visualization technology also plays a key role. Users need to interact with the representation and manipulates the structures, shapes and colors to reveal hidden properties.

B. Gephi Platform

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs[8]. It uses a special 3D render engine to render large graphs in real-time. Powered by its ad-hoc OpenGL engine, Gephi is pushing the envelope on how interactive and efficient network exploration can be.

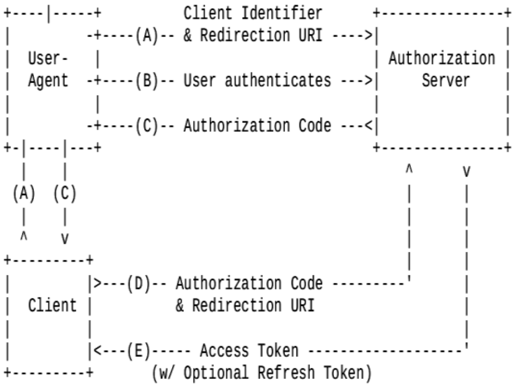
Layout algorithms give the shape to the graph. Gephi provides state-of-the-art algorithms layout algorithms, both for efficiency and quality. The Layout palette allows user to change layout settings while running, and therefore dramatically increase user feedback and experience.

Gephi is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. The goal of the Gephi platform is to help data analysts to intuitively discover patterns, make hypothesis, and isolate structure singularities or faults during the data analysis process.

III. DATA MINING SCHEME

A. Preparatory Work

We choose Sina Weibo as our research object as it is the most famous microblog service company in China. Tweets are collected from Sina Weibo website through the open API (Application Programming Interface) platform. The authentication process of API using the OAuth 2.0 protocol is somewhat cumbersome. Fig.2 reveals the authorization code flow. After building some web application in the API platform, the user agent may get the redirection URL and the client identifier, such as “app_key” and “app_secret” from the client. The authorization server then will send an authorization code to the user agent on the basis of them. After that, the client needs to transmit again the authorization code along with the redirection URL to the authorization server and it will response back with an access_token to the client. Finally, the client obtains OAuth 2.0 authorization of the Sina Weibo API.



```
+----|-----+          Client Identifier      +---------------+
|         -+----(A)-- & Redirection URI ---->|               |
|  User-   |                                 | Authorization |
|  Agent  -+----(B)-- User authenticates --->|     Server    |
|          |                                 |               |
|         -+----(C)-- Authorization Code ---<|               |
+-|----|---+                                 +---------------+
  |    |                                         ^      v
 (A)  (C)                                        |      |
  |    |                                         |      |
  ^    v                                         |      |
+---------+                                      |      |
|         |>---(D)-- Authorization Code ---------'      |
|  Client |          & Redirection URI                  |
|         |                                             |
|         |<---(E)----- Access Token -------------------'
+---------+       (w/ Optional Refresh Token)
```

Fig. 2. Authorization code flow

B. Propagation Path of the Tweets

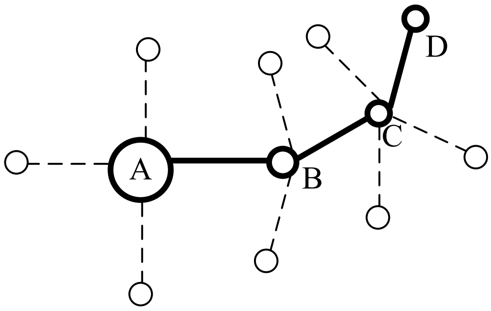
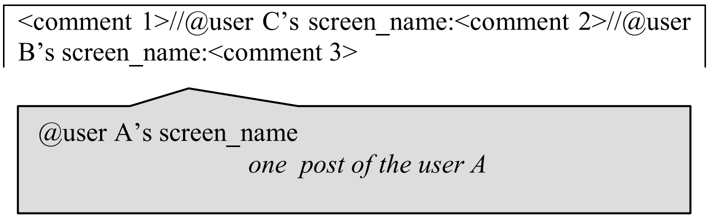


Fig. 3. The propagation path of one tweet

As shown in Fig.3, the bigger node A represents the original tweeter who creates the seed account and the linked nodes are the followers. A solid bold line means that the tweet has spread through the link by a retweet. After the user A posts a tweet, user B who is one of the four followers of the user A, may read this tweet and retweet it according to his own interests. And user C retweets it again who is the follower of the user B, and then the user D repeats the work. While the user retweets the message, he can comment it within 140 words. The system will display the words “retweeting the post” automatically if the user directly retweets without any comment in the Sina Weibo. Finally, in the homepage of the user D, it is shown as follows:



<comment 1>//@user C's screen_name:<comment 2>//@user B's screen_name:<comment 3>

@user A's screen_name

one post of the user A

Fig. 4. The display of the homepage of the user D

In Fig.4, the comments 1, 2, 3 respectively corresponding to the user D, C and B's comment are optional. If all the comments are default, the last comment 3 must show the words "retweeting the post". In Sina Weibo, "//@"is the tag indicating the retweet relation between a pair of nodes, so the path of this tweet diffusion creates a chain as shown in Fig.5.

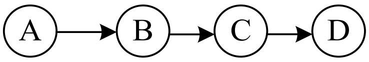


Fig. 5. The path of this tweet diffusion

The tweet is originally posted by user A, and is retweeted by user B, C and D in turn. User B is the first retweeter among them.

C. *Visualization Process of the Tweet Diffusion*

It is interesting that the visualization process of the tweet diffusion need to use the API in Sina Weibo.

Step 1: The integer ID of the objective tweet can be obtained through its URL. The ID parameter is passed as part of the URL itself.

Step 2: After the client accesses the API, several fields which are formatted JSON(JavaScript Object Notation) data such as reposts_count, comments_count, retweeted_status, user, geo, text, created_at, and so on, will be returned through the request parameter ID.

Step 3: The "text" field, such as:

"<comment 1>//@user C's screen_name:<comment 2>//@user B's screen_name:<comment 3>" in the above example, may be divided using the tag "//@" and results will be written to an CSV(Common Separated Value) file.

Step 4: The "screen_name" can be extracted from the CSV file and then nodes and directed edges between two nodes will be added into the Gexf(Graph Exchange XML Format) file.

Step 5: In the Gephi platform, the Gexf file needs to be loaded. It is necessary for the visualization process to select an appropriate layout algorithm and finally the visualization result of the tweet propagation will be exported to a PDF or PNG file.

IV. VISUALIZATION RESULTS

We find two typical tweet propagation path modes after researching several tweets propagation paths by means of the API in Sina Weibo. They are the dandelion mode and the double-star mode. Here are two typical representatives for the tweet spread path.

A. *The Dandelion Mode*

The URL of the representative tweet of the dandelion mode is http://www.weibo.com/1892680923/zdwwStWgM, which was created at 16:24 on January 8, 2013 and the data acquisition time was at 18:43 on May 29, 2013. The original author of this tweet is "Mathematical Culture" who is an adding V authenticated user. Fig.6 shows the diffusion path of the tweet. The tweet is retweeted the total 437 times. Repeat retweeting is regarded as only one edge, so there are actual 432 nodes and 435 edges in this transmission network graph. In the process of 437 times retweeting, there are 266 times completing only one-level spread and they are corresponding red nodes and edges in the graph. Blue nodes indicate that they participate in two-step flow or more spread. Gray edges mean links between the original node and all two-step flow nodes. Fig.6 presents the dandelion secondary explosive propagation pattern, which corresponds to the diffusion path characteristics of the grassroots celebrity microblog.

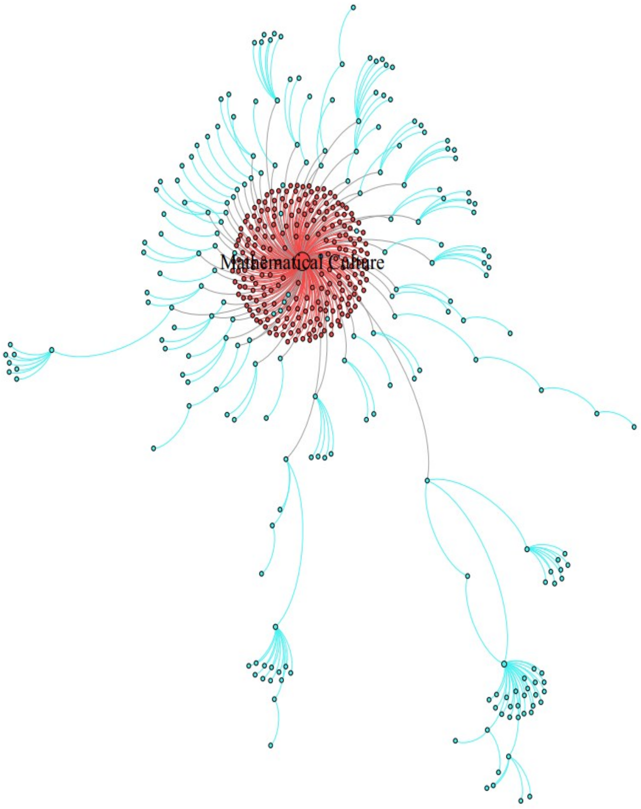


Fig. 6. Dandelion secondary explosive propagation pattern

B. *The Double-Star Mode*

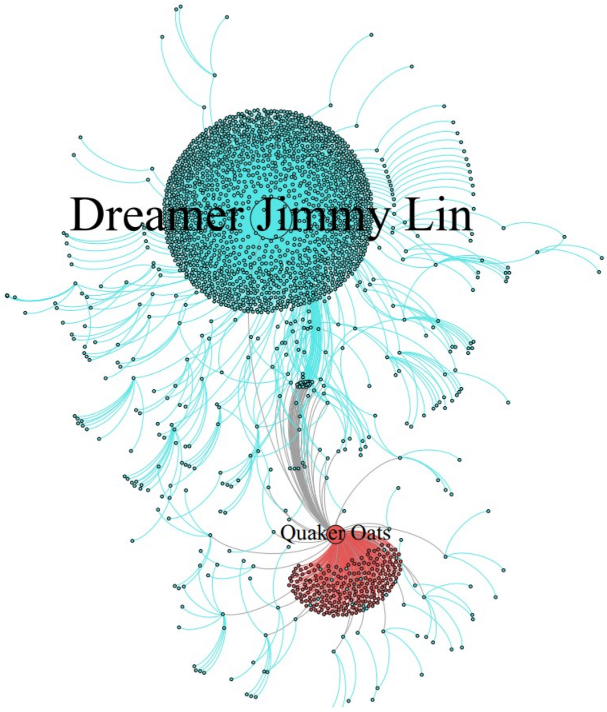


Fig. 7. Double-star explosive propagation pattern

The URL of the representative tweet of the double-star mode is http://weibo.com/1738475764/zvU3pC0vl, which was created at 13:23 on May 9, 2013 and the data acquisition time was at 11:13 on May 28, 2013. The original author of this tweet is "Quaker Oats" who is the official microblog account of the Quaker Oats Company and also an adding V authenticated user. The tweet itself is an endorsement advertisement by Mr. Jimmy Lin. Fig.7 shows the diffusion path of the tweet. The tweet has been retweeted the total 2353 times when the data is collected and there are actual 2223 nodes and 2314 edges in this transmission network graph. In the process of 2353 times retweeting, there are only 288 times completing just one-level spread. In fact, after Mr. Jimmy Lin retweeted the tweet at 9:32 on May 10, 2013, the tweet was retweeted many times by his large number of followers. As shown in Fig.7, the node "Dreamer Jimmy Lin" which is the microblog account of Mr. Jimmy Lin plays the most important retweeting role in the tweet diffusion process. The out-degree value of the node "Quaker Oats" is 299, but the node "Dreamer Jimmy Lin" is 1740. So the double-star explosive propagation pattern is formed, which corresponds to the diffusion path characteristics of the advertisement microblog with celebrity endorsements, such as movie stars, famous athletes, et al or the microblog controversy between both celebrities.

V. CONCLUSION

In the aim of better understanding the tweet diffusion process, the visualization method is studied.

From the visualization results of the tweet propagation path, we can easily answer the following two questions: 1) What is the role of this node in the tweet diffusion process? 2) Who is the key role in the tweet diffusion process? It is helpful for the analysis of the node's influence in the next work.

ACKNOWLEDGMENT

This paper is supported by the National High Technology Research and Development Program of China (863 Program) under Grant NO. 2013AA01A603.

REFERENCES

[1] W. Will, A.Stuart, and W.Roger, "Retweeting: a study of message-forwarding in twitter,'' Mobile and Onine Social Networks(MOSN), 2011 Workshop on Milan, pp. 13-18, Sept. 2011.

[2] Tatsuro Kawamoto, "A stochastic model of tweet diffusion on the twitter network,"Physica A,vol. 392, pp.3470-3475, 2013.

[3] Jose luis Ortega,Isidro F.Aguillo, "Network visualisation as a way to the web usage analysis," Aslib proceedings: New information perspectives, vol. 65, pp.40-53, 2013.

[4] Patrick M.Dudas, "Cooperative, dynamic twitter parsing and visualization for dark network analysis," iConference Proceedings, pp. 623-632, 2013.

[5] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonc¸alves, S. Patil, A. Flammini, and F. Menczer, "Truthy: mapping the spread of astroturf in microblog streams," InProceedings of the 20th international conference companion on World Wide Web, pp. 249–252, 2011.

[6] Hakan Kardes, Abdullah Sevincer, Mehmet Hadi Gunes, and Murat Yuksel, "Six Degrees of Separation among US Researchers," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, pp. 654-659, 2012.

[7] Donghao Ren, Xin Zhang, Zhenhuang Wang, Jing Li and Xiaoru Yuan, "WeiboEvents: A Crowd Sourcing Weibo Visual Analytic System," Proceedings of IEEE Pacific Visualization Symposium, Yokohama, Japan, pp.330-334, Mar. 2014.

[8] Bastian M, Heymann S, Jacomy M, "Gephi: An open source software for exploring and manipulating networks," International AAAI Conference on Weblogs and Social Media, 2009.