International Journal of Grid Distribution Computing
Vol.7, no.3 (2014), pp.199-204
http://dx.doi.org/10.14257/ijgdc.2014.7.3.20

A Study on Strength of Sina Weibo

Chen Fu[1], Zhan Shaobin[2] and Shi Guangjun[3]

[1]*Department of Computer Science and Technology, Beijing Foreign Studies University, Beijing 100089, China*
[2]*Shenzhen institute of information & technology, Shenzhen institute of information & technology, Guangdong, 518172, China*
[3]*Computer Network Information Center of Chinese Academy of Science, 100089, Beijing, China*

Abstract

With the widespread use of mobile smart device, and its ubiquitous network access capability, microblog have important influence. But to one node in microblog space, how to measure its importance is a key problem. In this paper, we propose a method to evaluate and compute the influence of one node in sina Weibo. An algorithm named MicroV is proposed to quantify the strength of one user in the microblog space. We use the algorithm to evaluate 60 thousand uses in the Sina Weibo.

Keywords: *online social network, MicroV, microblog space measure*

1. Introduction

With the developing of microblog, it has become a global social tool. The powerful information dissemination capabilities have triggered the curiosity of many scholars. And as the strong positive effect Twitter played in the U.S. presidential election of Obama administration, its social influence is steadily on the increase. The Sina Weibo have been gained considerable popularity and grown at an unprecedented rate in China. 140 characters or less convey to the people's livelihood concern for human interests, social justice, the problem of corruption, the battle of the value system, to verify the truth and the facts. The identity analysis of microblog as virtual society and its social management applications are imperative. After crawling large amounts of pages, we gathered much basic user information about microblog. Many characteristics of its users and contents as well as that of its opinion propagation were found after the information extraction and statistics.

With the increase of its influence about microblog such as sinaweb, many auxiliaries functions related to the one user can be found. For example, a wide variety of advertisements can be seen in sina Weibo space. And many celebrities may use it as a tool to disseminate the information. When we get the information from one user in microblog, we may evaluate the truthfulness according to the influence unconsciously. We seem to deem the words from the bigV, which have numerous fans. And this influence may be very important to person who wants to obtain a benefit. Advertisers, users and the content providers have different objectives. For the benefit from this virtual space influence, there are a lot of fake bigV. The fans can be added by software tools, so the amount of fans increase. And the cost to added fans is very low. The popularity of microblog greatly depends on the quality and integrity of contents and user. Also, cybercriminals have interest on the influence too. The fake influence can bring a disaster in sina Weibo which have more then 500 million users. More then 100 million messages are posted each day on Sina Weibo. Weibo has transformed from edge

Supported by the National Natural Science Foundation of China under Grant No.61170209, Shenzhen strategic emerging industry development funds(JCYJ20120821162230172), Program for New Century Excellent Talents in University No.NCET-13-0676.
ISSN: 2005-4262 IJGDC

media to the cutting edge of news media. Measuring the influence and understanding the behavior may be very important. So how to judge the user's fans becomes a very important problem. Instinctively, we can judge from many aspects by human being. But it may be very difficult, sometimes may be mission impossible. We can measure the real impact of a user in social networking in Sina Weibo with quantity methods. According to the influence computing, we can rank nodes based on strength of one nodes owns.

2. Related Word

We now live in a world of net connections. And the connections throw network play important role in our life. Many researches have pay attention to the topic related to the network connections. Node importance on Twitter is investigated in [1]. But its aim is "The centrality measures of a node importance do not show how important users are." And on the contrary, we want to show how important the node is.

The rest of this paper is organized as follows. Section 2 describes the methodology for data collection. The model of computing the strength of the Sina Weibo nodes is presented in Section 3. In Section 4, we propose measurement of the node influence in sina Weibo community. In Section 5 we study the law of followings dynamic evolution. Law of fans dynamic evolution and law of sina Weibo evolution are studied in Section 6 and 7.The paper is concluded in Section 8.

3. Sina Weibo Data Collection

To show the trend of influence strength, we compute 5 million nodes. So we have to collect the sinaWeibo data firstly. We have collect 5 millions nodes personal information, include the username, username, screen name, follower numbers et al.

(1) Data Collection With SINA API

To collect with SINA API, we have to register an App named Smile-Face. After the configuration, we can gain the two values named App Key，App Secret. With the callback address, client_ID, client_SERCRET, access_token, we get the data related to one node. But after some experiment, we can see this method will not work if we want to get huge dataset. For example, after getting some data the program have to wait a period. So if we want to get millions dataset, we have to with agent to collect data automatically. In some special condition it also can work. For example if we can develop a program based on the SINA API to provide some service such as who follow me, and I follow who with a program.

(2) Data Collection With Crawler

To avoid above limitation, we develop a crawler named Rosetta in Python with identity information following the client-server model to collect the user profiles information, which is depicted in Figure 1.

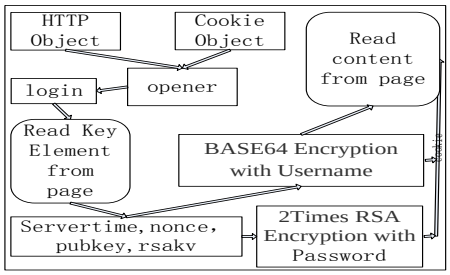

Figure 1. Crawler Named Rosetta

Although we can collect information related to one node in Sina Weibo environment, it is almost impossible to collect all nodes within a short time. But we do think it is critical to master the core technology to collect the data in micro bolging space. Especially with the emerging of big data science, the dataset is the foundational basis of all kinds the data related application. To simulate the people logining, we analysis the interactive procedure of HTTP with Http Fox. Of course, the URL may be changed with time going on, but we can analysis the HTTP protocol with the tools such as Http Fox with the change. Parallelism with multithreading, multiprocessor is considered.

Finally, with above methods, wed also download the data from the community resource, we collect more the 5 million the Sina Weibo node profile information. To keep the dataset probability distribution is in uniform distribution, we focus on a special set whose follower number more then 1 million.

4. Measurement of The Node Influence in Sina Weibo Community

In this section we focus on the method of the influence to one node in Sina Weibo community. There have some works to this problem. Generally, the influence may be decided by many factors, such as the meaning of words related to one node, the number of comments, the affiliation, and the number of forwarding. But with the consideration of simpleness, we only consider the number of follower. Just because we only want to find fake 'BigV' and give the trend of Sina weib community.

(1) Definition

Followings: the number of one node following the other nodes.

Fans: the number of one node is followed by others.

Weibo Num: the message the node publishes, the 'tweets' in Sina Weibo.

(2) Data Statue

We have collected the data with initial process. According to this paper's topic, we have collected about 5 millions nodes information and the this will be continued. Also, we have collected about 13G content related to about 1 thousand nodes for further research. To one node, the ID, screen name, Sex, VIP description, Followings, Fans, *et al.*,.

5. Law of Followings Dynamic Evolution

(1) Power Law Followings Number Less Then 50

It is impossible to collect all the data from Sina Weibo space. According to our data, we have founded several interesting patterns. We select randomly from the node information and the node number nodes with number followings, as show Figure 1.

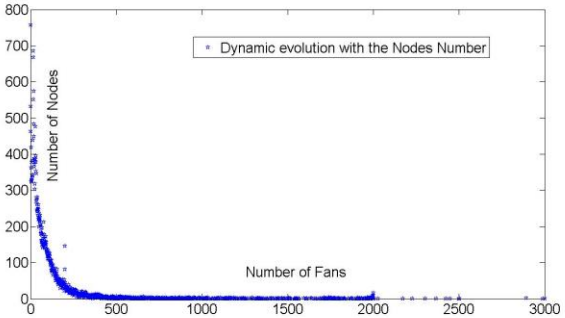

Figure 2. Followings Number Dynamic evolutions

Just as many references show, intuitively the number of nodes with number followings will follow the Power law. So we show the same set data with double log Figure, just as showed Figure 3.

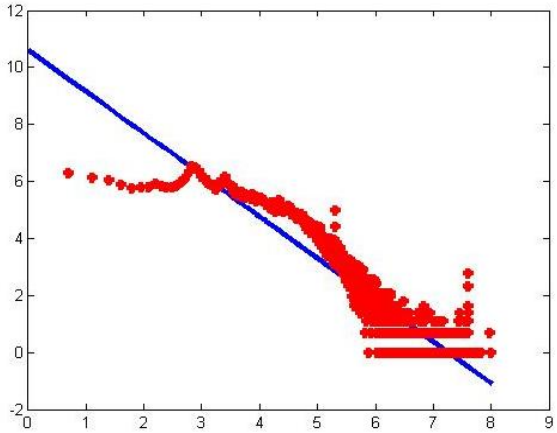

Figure 3. Double Log

From the Figure 3 we can get that the number of followings don't follow the power law strictly. But there are still have several interesting law. Firstly, the number less then 150 follows power law strictly. This can be seen from Figure 4.

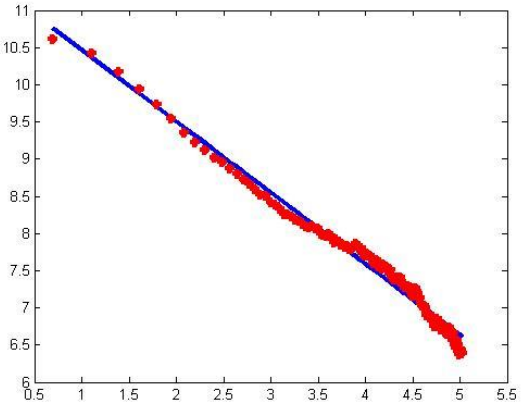

Figure 4. Followings less then 150

The power law linear function is:

Y=-0.9578x+11.42 …… (1)

(2) ' 2Lines' Law Followings Number Less then 2000

The second law we named '2 Lines', that is followings between 50 to 90, between 500 to 3000 follow linear strictly, but with different slop. The '2line' law is showed in Figure5.

The linear equations are as follows:

y=-18.92x+4201 ……(2)

y=-0.01385x+32.09……(3)

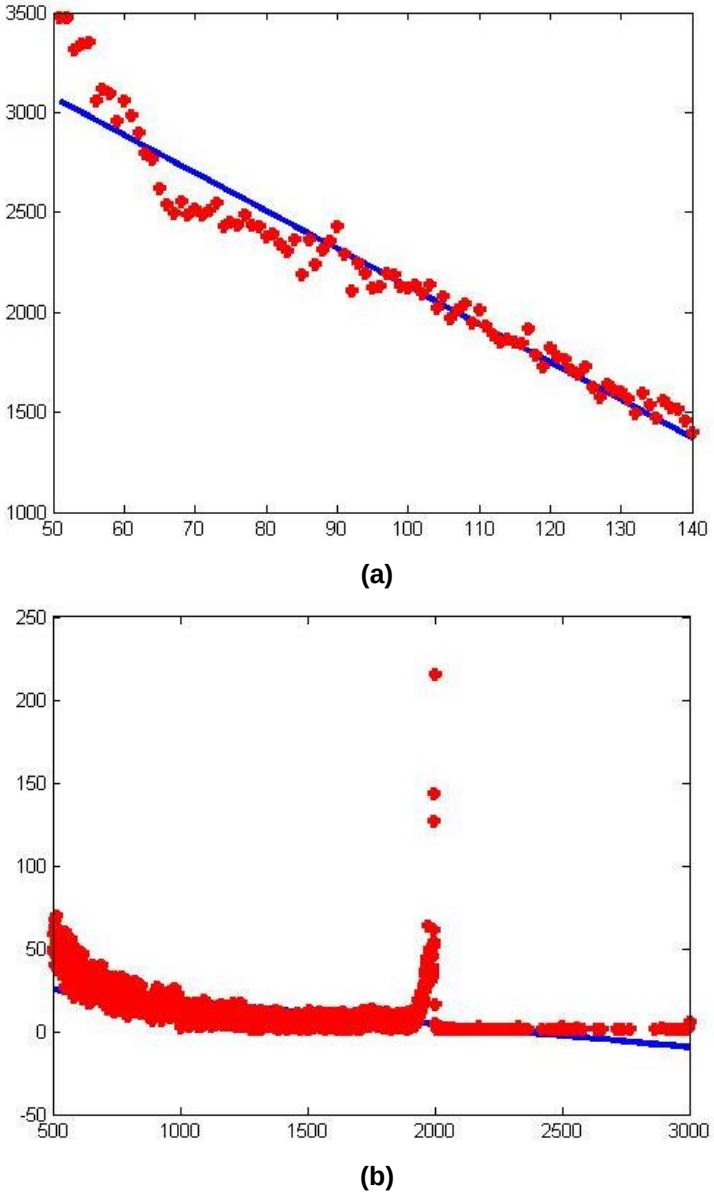

Figure 5. '2 Line' Law

(3) Followings Number 'Sina2000' and 'Sina500'

There are ranks of fans number, membership less then '2000', senior membership nodes less then 2500, and extremum limitation less then 3000. But we found close to number 2000, the number increases dramatically. The other one interesting thing is, just as the Dunbar's number 150, in Sina Weibo is 500. We all know that 140 Characters or Less is allowed in micro blogging, which because of the mobile Tel in U.S is limited to 140. And many may knows Dunbar's number 150, which is a suggested cognitive limit to the number of people with whom one can maintain stable social relationships. We have found a very interesting fact that number near '2000' in Sina micro blogging, which mean many special features. The followings number more then 2000 is dramatically decrease, which is showed in Figure 6. We have crawl and test it for many times. We also can draw that before 2000, the number varies

smoothly. But just as the followings numbers there are often have a strange high value and then maintain a very low value.

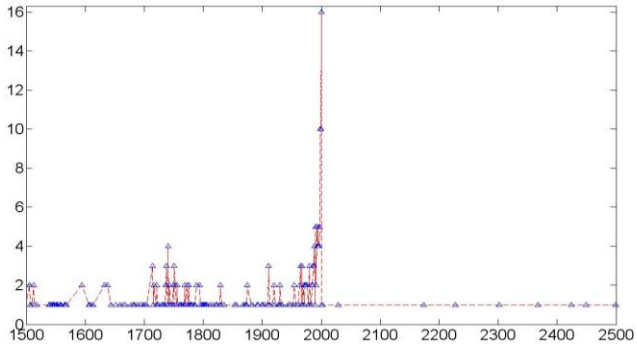

Figure 6. Special Followings Number '2000'

6. Conclusion and Future Work

In this work, we propose the methods of the Sina Weibo data collecting and regular evolution pattern followings, fans, and Sina Weibo. Some phenomenon is very interesting. Such as near limitation 2000, the number increase dramatically. Also these nodes publish large number Weibo. It may be the commercial nodes for advertisement publish. High fans don't necessary mean large Weibo. The reason may that the large number fans come from the real society influence but not from the Weibo publish. And the quality and quantity of the Weibo have less important then influence in real society. By studying part of followings and fans, we draw some statistical differences and law. We also found linear relations between certain range and power-law relations in some other range.

In our future work, we will analyze special point nodes with classification. Digging other potentially useful features, and more importantly, propose more effective models for nodes classification and nodes recommend.

References

[1] X. XiaoBing, Z. Gang, H. YongZhong, C. HaiYong and X. Ke, "Dynamic evolution of collective emotions in social networks a case study of Sina Weibo", Science China, vol. 56, **(2013).**
[2] H. Nguyen and R. Zheng, "A Data-driven Study of Influences in Twitter Communities", CoRR abs/1307, vol. 4264.
[3] V. Lampos, T. Lansdall-Welfare, R. Araya and N. Cristianini, "Analyzing Mood Patterns in the United Kingdom through Twitter Content", CoRR abs/1304, vol. 5507, **(2013).**
[4] J. Fabrega and P. Paredes, "Social Contagion and Cascade Behaviors on Twitter", Information, vol. 4, no. 2, **(2013),** pp. 171-181.
[5] R. Lee, S. Wakamiya and K. Sumiya, "Urban area characterization based on crowd behavioral lifelogs over Twitter", Personal and Ubiquitous Computing, vol. 17, Issue 4, **(2013)** April, pp 605-620.
[6] J. Fabrega and P. Paredes, "Social Contagion and Cascade Behaviors on Twitter", Information, vol. 4, no. 2, **(2013),** pp. 171-181.
[7] L. M. Aiello, G. Petkos, C. Martin and D, Corney, "Yahoo Res., Barcelona pain, Sensing Trending Topics in Twitter", IEEE Transactions on, Multimedia, vol. 15, Issue 6, pp. 1268-1282.
[8] S. Cheng, H. Shen, J. Huang, G. Zhang and X. Cheng, "StaticGreedy: solving the scalability-accuracy dilemma in influence maximization", Proceedings of the 22nd ACM international conference on Conference on information & knowledge management, **(2013)** October 27-November 01, San Francisco, California, USA.

