KSS'2013 NINGBO

Proceedings

Knowledge Creation Towards Emergency Management

Shouyang Wang, Yoshiteru Nakamori and Weiliang Jin (eds.)

JAIST Press

ISBN: 978-4-903092-36-2

A Preliminary Research of Pattern of Users' Behavior Based on Tianya Forum

Yongliang Zhao[1], Xijin Tang[2]
Academy of Mathematics and Systems Science,
Chinese Academy of Sciences,
Beijing 100190, China
[1]zhaoyongliang12@mails.ucas.ac.cn
[2]xjtang@amss.ac.cn

Abstract

The fast development of Internet makes it an ocean of data. In order to find useful information, the Web mining technology comes into being. We get data from Tianya Forum with spider program and store them in file system and database; then we investigate the pattern of users' behavior based on the data of "Tianya Zatan" board in 2012. The result is that users spend less time online on non-working days, the distribution of clicks is described by the mixture of Poisson distribution and power law, while the distribution of replies is power law. What is more, we evaluate the hot degree of posts with four different methods and push daily ranking lists to study the pattern of users' behavior in hot posts.

Keywords: Web mining, users' behavior, Poisson distribution, power law, hot degree

1 Introduction

With the fast development of Internet, it has become an ocean of data. Internet data are typical big data, which have the characteristics of universality, complexity, etc[1]. It has become a research hot spot in data mining, business search, etc. to find useful contents and patterns from Internet data. Web mining includes Web content mining, Web structure mining and users' behavior mining, of which users' behavior mining is of great importance. Zhang, Tang and Yoshida introduced an approach to Web information processing and applied to mining of one famous scientific forum in China based on Web text mining[2]. Ari, Alison and Kate, *et al.* used "Google Trends" for epidemiological research, which is one of the earliest cases of users' behavior mining[3]. Fabio, Barbara and Luca, *et al*. provided a quantitative analysis of users' behavior in Friendfeed[4]. Li did an empirical study on users' behavior in hot Weibo at her master thesis[5]. Cui, He and Liu discussed hot issues and evolutionary trends in online forums[6].

So far, Tianya Club, with more than 80 million registered users, covers more than 200 million users every month[7]. It is a leading focused platform for Internet events and Internet celebrities in China. Then it is useful to study the pattern of users' behavior in Tianya Forum. For example, by studying the pattern of users' behavior, we have a better understanding of online behavior and know the development and evolution of Internet events. It also helps steer the development of Internet events along the positive line and build the harmonious society.

The rest of the paper is organized as follows: in Section 2 we introduce the data processing of Tianya Forum, including data acquisition and data storage. In Section 3 we study the pattern of users' behavior based on the data of "Tianya Zatan" board in 2012, and discuss the impact of non-working days, which include holidays and weekends, on users' behavior and the distribution of clicks and replies. In Section 4 we compare to evaluate the hot degree of posts with four different methods and push daily ranking lists to who is concerned. Section 5 is our concluding remarks.

2 Data Acquisition and Storage of Tianya Forum

We use Web mining technology to mine Tianya Forum. The data processing of Tianya Forum is as shown in Figure 1. Data acquisition and storage goes on automatically every day.

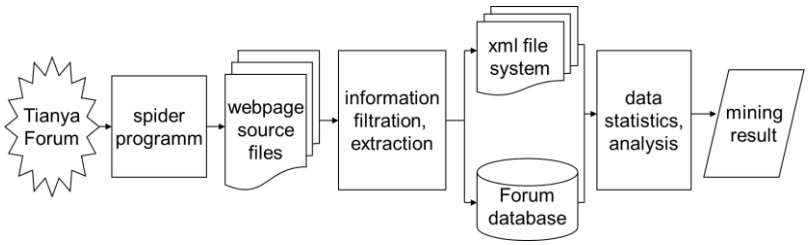

Figure 1. The Data Processing of Tianya Forum

2.1 Data Acquisition

The data acquisition of Tianya Forum is as shown in Figure 1. We use spider program to download data from Tianya Forum every day. First, the spider program collects updated information of "Tianya Zatan" board and stores it on local machines. Then we get posts' contents and their updated information from source webpages after webpage filtration, feature extraction and information extraction, and put them into database and file system[8][9].

There are two key problems in data acquisition. The first is that Tianya Forum keeps changing aperiodically, such as launching new versions and founding new boards. So we have to adjust our spider program. The second is that the spider program sometimes did not work well, leading to data loss. So the missing data have to be re-crawled to keep data completed.

2.2 Data Storage

The data storage of Tianya Forum adopts both mysql database and xml file system so as to have both advantages. For the first way, which uses mysql database, it contains two tables: *post* and *postupdate*. The structures of the two data tables are as follows:

post(pID, title, author, content, link, date_posted, time_posted, boardID)
postupdate(pID, date_update, time_update, hits, replies)

where *post* table stores posts' information. "pID" is the auto-increment primary key, which identifies a post uniquely. Other fields denote title, author, content, source webpage address, posted date, posted time and board of the post respectively. *Postupdate* table stores posts' updated information. Both "pID" and "date_update" are the composite keys. Other fields respectively represent the primary key, updated date and time, clicks and replies of a post.

For the second way, which uses xml file system, the contents are the same as mysql database. But xml file system is standard for data management and exchange. Combining both ways can prevent data loss and damage.

3 Analysis of Tianya Users' Behavior

For Tianya Forum, individual user's behavior mainly includes posting, click and reply. Posting represents that users release posts and want to be concerned. Click means that users read and are interested in the posts. Reply represents that users join the discussions. These behaviors are described by quantitative data at macro level, such as the number of posts, clicks and replies. This paper takes them as the index of users' behavior when studying the pattern of users' behavior. Detailed research includes statistics of posts, the impact of non-working days, which include holidays and weekends, and the distribution of clicks and replies.

Tianya Forum has many boards, such as "Tianya Zatan" and "People's Voice", of which "Tianya Zatan" board is active and highly related to Internet events and public opinion. So this paper uses the data of "Tianya Zatan" board in 2012. The data include both new posts and updated posts daily. New posts refer to the posts that are published on that very day, while updated posts refer to the posts that are updated on that very day, including new posts on the same day.

3.1 Statistics of Posts

For the data of "Tianya Zatan" board in 2012, we count both new posts and updated posts daily. Due to unexpected reasons, some data are incomplete and lost. For the incomplete data, about 2.73% of the total data, we re-crawl them; for the missing data, about 1.64% of the total data,

we use the average of the amount of daily posts in that very month instead. After data pre-processing, we find that the amount of new posts is 409,717, about 1,119 per day, the daily maximum is 1,927, while the daily minimum is 185; that of updated posts reaches 1,195,125, about 3,265 per day, the daily maximum is 4,755, while the daily minimum is 789. The numbers of monthly new posts and updated posts are as shown in Figure 2.

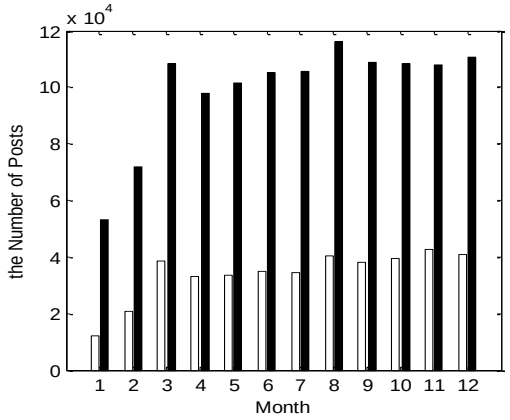

Figure 2. Statistics of Monthly Posts

From Figure 2, we see that the ratio of monthly new posts and updated posts is about 1:3. The month average of new posts is 37,658.4, and the average of updated posts is 107,004.4 from March to the end of 2012, which indicates that both new posts and updated posts keep stable. It shows that users' behavior keeps stable in "Tianya Zatan" board.

3.2 The Impact of Non-working Days

Non-working days include holidays and weekends, we consider them separately.

1) Holidays

In order to study the impact of holidays on users' behavior, we define Posting Rate of holidays as shown in Eq. (1).

$$PR_h = \frac{average\ number\ of\ daily\ posts\ on\ Holidays}{average\ number\ of\ daily\ posts\ before\ and\ after\ m\ days\ of\ Holidays} \quad (1)$$

where m is the length of holidays. PR_h reflects the percentage of the number of posts on holidays.

We study the impact of official holidays on users' behavior, using PR_h. The holidays in 2012 that are taken into consideration include New Year's Day (Jan. 1-Jan. 3), the Spring Festival (Jan. 22-Jan. 28), the Qingming Festival (Apr. 2-Apr. 4), May Day (Apr. 29-May 1), the Mid-autumn Festival and National Day (Sept. 30-Oct. 7). As the Mid-autumn Festival and National Day are linked together in 2012, we see them as one holiday. The values of PR_h of those holidays are listed in Table 1.

From Table 1, we find that no matter for new posts or updated posts, all their PR_h are below 1, and the minimum reaches 0.58. It shows that holidays have a great impact on users' behavior, and the total posts on holidays are about 2/3 of workdays.

2) Weekends

In order to study the impact of weekends on users' behavior, we define Posting Rate of weekends as shown in Eq. (2) in the same way.

$$PR_w = \frac{average\ number\ of\ daily\ posts\ on\ weekends}{average\ number\ of\ daily\ posts\ on\ workdays} \quad (2)$$

PR_w reflects the percentage of the amount of posts on weekends. We study the impact of weekends on users' behavior using PR_w. First, we rank all of the number of daily posts in 2012 and compute PR_w by week. Then we get the average of all the weeks. There are three scenarios to compute the average: Scenario 1 uses all the weeks, Scenario 2 removes the weeks that are related with holidays, Scenario 3 removes all the weeks before or after one week of holidays. The averages are listed in Table 2.

Table 1. Posting Rate of Holidays in 2012

	New Year	Spring Festival	Qingming Festival	May Day	National Day	Average
New Posts	0.68	0.62	0.81	0.66	0.59	0.67
Updated Posts	0.58	0.71	0.94	0.83	0.76	0.76

Table 2. Posting Rate of Weekends in 2012

	Scenario 1	Scenario 2	Scenario 3	Average
New posts	0.83	0.80	0.75	0.79
Updated Posts	0.90	0.89	0.86	0.88

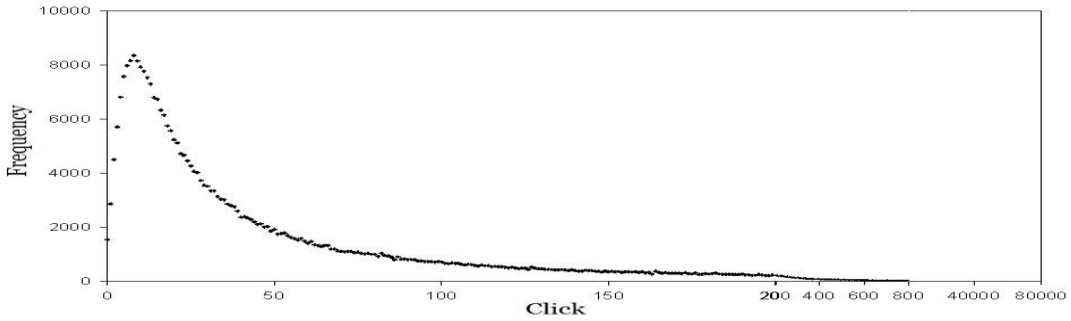

Figure 3. The Distribution of Clicks

From Table 2, we find that no matter for new posts or updated posts, all of the averages of PR_w are below 1, and the minimum reaches 0.75. As Scenario 3 removes the impact of holidays completely, it represents the impact of weekends on users' behavior in the most accurate way. The result shows that weekends have a great impact on users' behavior, and the number of posts on weekends is about 4/5 of workdays.

Obviously, non-working days have a great impact on users' behavior. The reason why the Posting Rate of non-working days is low is possibly that most of Internet users are at rest or do something else in the real world and spend less time online correspondingly. This phenomenon can also be found in Weibo[10].

3.3 The Distribution of Clicks

Clicks reflect whether users are interested in some posts. In order to study the pattern of clicks, we count new posts that have the same clicks based on the data of "Tianya Zatan" board in 2012. The statistics result is as shown in Figure 3.

From Figure 3, we see that when the clicks of posts increase from 0 to 8, the amount of posts rises rapidly; when clicks get to 8, the amount reaches the maximum, getting 8,351; when clicks are more than 8, the number goes down with the increasing of clicks, and the slope of the curve slows down. Posts whose clicks range between 0 and 130 make up 80% of all the posts. The amount of the posts, whose clicks are more than 2000, is less than 10, covering 1.64% of the total amount. It tells us that only a very few posts have very high clicks, while the clicks of most posts are very low. That is to say, users only have interest in a very few hot posts, while most posts are not cared much.

In order to verify the distribution of clicks, we take the logarithm for both clicks and the corresponding number of posts (log10). Then we apply simple linear regression to them using ordinary least squares. The result is presented in Figure 4, and the regression equation is as shown in Eq. (3).

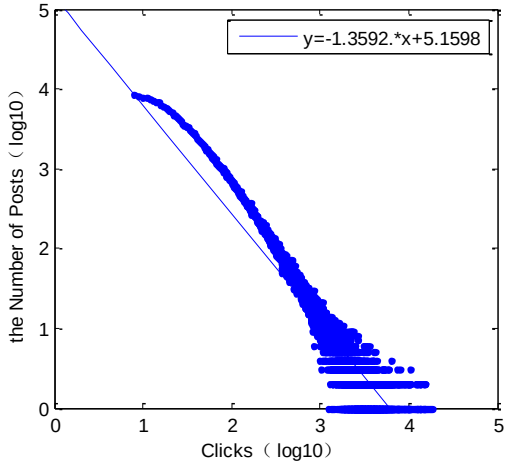

Figure 4. The Distribution of Clicks.

$$y = -1.36 * x + 5.16 \qquad (3)$$

where the goodness of fit is 0.86; The result of F test is 34185.01, far greater than threshold 3.84; the result of t test is 184.89, far greater than threshold 1.96. It shows that the regression is linear, which thereby verifies the distribution of clicks is described by power law when clicks are more than 15.

More study shows that the distribution of clicks is the mixture of both Poisson distribution and power law. When clicks are lower than 15, its distribution follows Poisson distribution, of which the mean is 8.3. When clicks are more than 15, it is described by power law, which has been verified above.

The distribution of clicks shows that when the posts are just released, users click them out of interest, leading to a random action as the interest of users is different. However, when clicks reach a certain value, users choose to click hot posts, that is to say, users become onlookers. It is possible that this mechanism leads to the mixture distribution.

3.4 The Distribution of Replies

Replies show users' interest to further discuss some posts. In order to study the pattern of replies, we count new posts that have the same replies based on the data of "Tianya Zatan" board in 2012. Replies have only 692 discrete values with range of [0, 4984] except for an outlier 51960. The range of number of posts is [1, 163714]. The amount of posts with no replies reaches 163714, about 1/3 of all the posts. The statistics result is as shown in Figure 5.

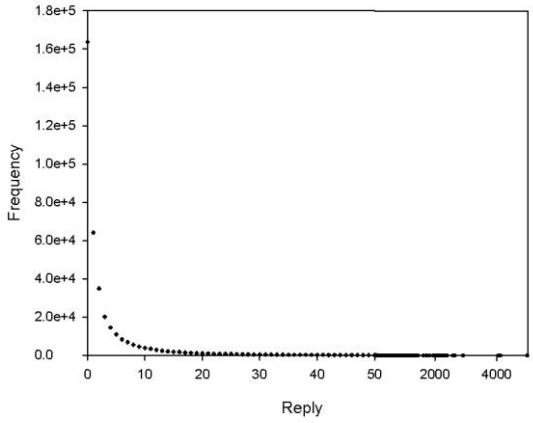

Figure 5. The Distribution of Replies

From Figure 5, we see that the amount of posts goes down with the increasing of replies of posts, and the slope of the curve slows down. Posts whose replies are between 0 and 5 make up 80% of all the posts. The amount of posts, whose replies are more than 300, is less than 10, covering 0.17% of the total amount. It tells us that only a very few posts have very high replies, while most posts are of very low replies. Moreover, users only have interest in a very few hot posts, while most posts are not browsed.

In order to prove the distribution of replies is power law, we first take the logarithm for both replies and the corresponding number of posts (log10), then apply simple linear regression to them using ordinary least squares. The result is as shown in Figure 6, and Eq. (4) is the regression equation.

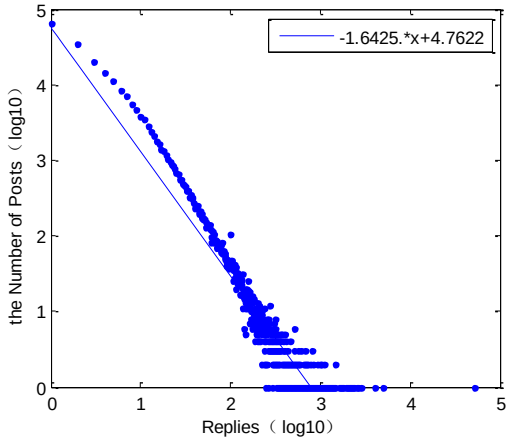

Figure 6. The Distribution of Replies.

$$y = -1.64 * x + 4.76 \qquad (4)$$

where the goodness of fit is 0.87; The result of F test is 4558.25, far greater than threshold 3.84; the result of t test is 67.51, far greater than threshold 1.96. It shows that the regression is linear, which thereby proves the distribution of replies is described by power law[11].

Power law is one of the most widely used distributions to describe human social behavior. For a post, except for "Online Water Army", which refers to the hordes of people out there that are paid to post comments on the Internet, only users who have read and think about it will reply and join the discussion. It possibly accounts for the reason why the distribution of replies is power law.

4 Users' Behavior in Hot Posts

In order to study the pattern of users' behavior in hot posts, we use four different methods to evaluate the hot degree of posts and push the ranking lists to who is concerned every day.

4.1 Hot Degree

The factors that have an impact on the hot degree of posts include clicks, replies, the ratio of replies and clicks, etc. This paper evaluates the hot degree of posts by the four methods as shown in Eq. (5-8).

No.	Title	Click	Reply	Ratio	Mix
1	南方日报杨兴乐记者受贿的文章太失水准(转载)(转载)	124787	0	0.00	56.39
2	网曝唐山远大职工变"临时工" "烧"完十亿资产玩破产(转载)	35782	175	0.00	25.95
3	政府大门清晨被砸，办公室惊现植物人	22478	7	0.00	10.55
4	吉林中院法官诈骗窝案"牵出"司法界"惊天假案"！(转载)	18970	52	0.00	11.48
⋮	⋮	⋮	⋮	⋮	⋮

Figure 7. Part of The Ranking List for New posts On August 17, 2013

No.	Title	Click	Reply	Ratio	Mix
1	深度解析"韩寒挑战方舟子"一战究竟谁赢了？（技术帖直播）	20015465	1409585	0.07	329.30
2	从噩梦到天堂：离婚四年的成长史（性、爱、事业及其他）连载	16319489	239858	0.01	87.63
3	水泊梁山那些基情燃烧的岁月:妖言水浒之大宋盛世(笑死算自杀)1166页更新	12270456	127435	0.01	55.37
⋮	⋮	⋮	⋮	⋮	⋮

Figure 8. Part of The Ranking List for Updated Posts On August 17, 2013

$$HD_i = h_i \quad (5)$$

$$HD_i = r_i \quad (6)$$

$$HD_i = r_i / h_i \quad (7)$$

$$HD_i = w_1 h_i / avg(h) + w_2 r_i / avg(r) + w_3 (r_i / h_i) / avg(r/h) \quad (8)$$

where i represents the i-th post; h_i is the clicks of the i-th post; r_i is the replies of the i-th post; $w_j (j = 1,2,3)$ is the weight and $\sum_{j=1}^{3} w_j = 1$, avg represents the average of all the posts that are taken into consideration.

These four methods measure the hot degree of posts from four different perspectives. Eq. (5) uses clicks as the hot degree of posts. But it cannot exclude some posts which have attractive titles but meaningless contents. Eq. (6) takes replies as the hot degree of posts. However, it misses some posts that are banned. Eq. (7) tries to find those posts with high clicks and replies. Eq. (8) weighs the factors of clicks, replies and the ratio by comprehensive consideration. We use the average of the ratio of replies and clicks instead of the maximum in ref. [6].

4.2 Daily Ranking Lists on Hot Degree

We write a program to push daily ranking lists to who is concerned based on hot degree. The program first ranks the daily posts by the hot degree, then extracts the top 20 posts and pushes them to who is concerned. The daily ranking lists include the ranking list for daily new posts and that for daily updated posts. As there are four methods to evaluate the hot degree of posts, each of the two ranking lists is divided into four parts, which are the ranking list of clicks, replies, the ratio of replies and clicks, and hot degree respectively. Figure 7 is a schematic diagram of the ranking list of clicks for new posts on August 17, 2013, while Figure 8 is for updated posts.

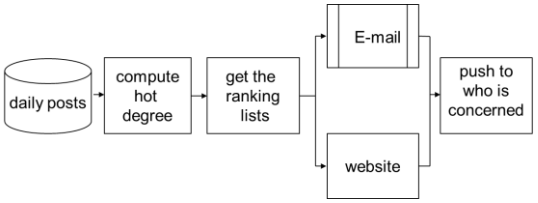

Figure 9. The Work Flow of The Program to Push Daily Ranking Lists

The work flow of the program to push daily ranking lists is as shown in Figure 9. It gets daily new posts and updated posts from mysql database, which is introduced in Section 2. Then the program computes the hot degree of the posts and gets the ranking lists. The program provides two ways to see the ranking lists. The first way is that the program automatically pushes them to who is concerned through E-mail every day, while the second way is that they can be visited through website. Interestingly, the ranking list for updated posts almost remains constant, which shows those very hot posts keep updated every day.

5 Conclusions

The rapid development of Internet technology has accumulated a large number of users' online data. Detecting the pattern of group behavior, mining perspective and perceiving organizational structure from Internet data has important theoretical and practical significance. It is useful

for grasping users' interest and the pattern of users' behavior, and collecting and analysing public opinion.

This paper studies users' behavior based on the data of "Tianya Zatan" board in 2012. The data are acquired by spider program and stored in file system and database. It shows that users have less time online on non-working days, which may give a hint when to post if we want our posts to be concerned with. From the distribution of clicks and replies, we know that posts whose clicks are more than 130 only account for 20% of all the posts, while posts whose replies are more than 5 only consist of 20% of all the posts. It tells us that users only show interest toward a very few hot posts. This paper also evaluates the hot degree of posts and ranks the daily new posts and updated posts to further study the pattern of users' behavior in hot posts, which are firmly related with Internet events and public opinion.

This paper has much space for improvement. First, this paper is just a preliminary study of users' behavior, more work needs to be done. For example, the pattern in updated posts, the relation between the distribution of clicks and replies, all of this is going to be studied. Second, the posts' contents need to be mined with users' behavior. Cao and Tang explored to extract topics from posts with LDA algorithm[12]. However, more analysis of contents needs to be done. Third, more reasonable evaluation of hot degree is to be explored and the program designed to push daily ranking lists needs to be improved.

Acknowledgments

This research was supported by National Basic Research Program of China under Grant No. 2010CB731405, National Natural Science Foundation of China under Grant No.711771187.

References

[1] V M Schonberger, et al. Big Data: A Revolution That Will Transform How We Live, Work, and Think (Chinese version, translated by T Zhou, et al). Zhejiang People's Publishing House, 2013.

[2] W Zhang, X J Tang, Y Taketoshi. AIS: Anapproach to Web information processing based on Web text mining. Systems Engineering-Theory & Practice. 2010, 30(1): 96-99. (in Chinese)

[3] A Seifter, A Schwarzwalder, K Geis, et al. The utility of "Google Trends" for epidemiological. Geospatial Health. 2010, 4(2): 135-137.

[4] F Celli, M Magnani, B Pacelli, et al. Social Network Data and Practices: the case of Friendfeed. Advances in Social Computing, Springer, 2010, LNCS Vol. 6007, pp. 346-353.

[5] Y Li. An Empirical Study on Users' Behavior and Information Dissemination in Hot Weibos. Master thesis, University of Chinese Academy of Science, 2013, Beijing. (in Chinese)

[6] L J Cui, H He and W Liu. Research on Hot Issues and Evolutionary Trends in Network Forums. International Journal of u- and e-Service, Science and Technology. 2013, 6(2): 89-98.

[7] http://help.tianya.cn/about/history/2011/06/02/166666.shtml.

[8] Z D Zhang. Design and Implementation of Tianya Forum Vision1.0-A Web Mining System Based on Tianya Forum. Master thesis, Graduate University of Chinese Academy of Sciences, 2012, Beijing. (in Chinese)

[9] Z D Zhang, X J Tang. A Preliminary Study of Web Mining for Tianya Forum. Proceedings of the 11th Youth Conference on Systems Science and Management Science, Wuhan: Wuhan University of Science and Engineering Press, 2011, 199-204. (in Chinese)

[10] Z Gao, Z Li, H Tu, et al. Characterizing User Behavior in Weibo. Mobile, Ubiquitous, and Intelligent Computing (MUSIC), 2012 Third FTRA International Conference on. IEEE, 2012: 60-65.

[11] W Cheng, H Zhong, J H Sun. Research on complexity of posts in network forums. Journal of Systems Engineering. 2009, 24(4): 385-391. (in Chinese)

[12] L N Cao, X J Tang. Prevailing Trends Detection of Public Opinions Based on Tianya Forum. The 14th International Conference on Intelligent Data Engineering and Automated Learning, Springer, 2013, LNCS Vol. 8206, 187-194.