Association for Information Systems
AIS Electronic Library (AISeL)

WHICEB 2015 Proceedings | Wuhan International Conference on e-Business

Summer 6-19-2015

Analyze the Trend of Post Replies Based on Linear Regression Model-----take Tianyawebsiteas examples

Jingwen Chen
School of Business Administration,Zhongnan University of Economics and Law economics, Wuhan,400000,China,
1973887130@qq.com

Shumeng Liao
School of Business Administration,Zhongnan University of Economics and Law economics, Wuhan,400000,China,
lsm940425@163.com

Yuan Yin
School of Business Administration,Zhongnan University of Economics and Law economics, Wuhan,400000,China

Follow this and additional works at: http://aisel.aisnet.org/whiceb2015

Recommended Citation
Chen, Jingwen; Liao, Shumeng; and Yin, Yuan, "Analyze the Trend of Post Replies Based on Linear Regression Model-----take Tianyawebsiteas examples" (2015). *WHICEB 2015 Proceedings*. 68.
http://aisel.aisnet.org/whiceb2015/68

This material is brought to you by the Wuhan International Conference on e-Business at AIS Electronic Library (AISeL). It has been accepted for inclusion in WHICEB 2015 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Analyze the Trend of Post Replies Based on Linear Regression Model-----take Tianyawebsiteas examples

Jingwen Chen[1], Shumeng Liao[2], Yuan Yin[3]*

[1]School of Business Administration,Zhongnan University of Economics and Law economics, Wuhan,400000,China

Abstract: In recent years, users spend more time on surfing the social networking than ever before. How to make the information spread rapidlywhen facing vast amounts of information?Scholars have conducted information dissemination in social networking. On the basis of previous research, the authors divide posts into popular posts and ordinary posts and then use the linear regression model to predict the replies at specified time. After comparing the difference between two types of posts, the authors concludethat ordinary posts could become popular posts if the posts could maintain a large number of replies within former five hours and increase replies by making use of community mechanism. This conclusion provides a reasonable proposal for enterprises and administrators to identify andrecommendpopular posts.

Keywords: information dissemination, linear regression model, information prediction

1. INTRODUCTION

With the rapid development of Internet technology, the well-known Facebook, Twitter and virtualcommunityhave become a place for people to obtain and share information. In Web2.0 mode, companies are inclined to use the Internet to promoteproducts and shape the corporate brand image by submitting favorable remarks. So it is essential for enterprises to control the popular news.

However, not all the information will attract sufficient attention. Some contentsgather user's attention, while some others do not. When popular information is recommended to the home page, the phenomenon of "richer get richer" would attract more users' attention.Most popular information comes from active social networking sitesbutthe authenticity needs review. Due to imperfect community mechanisms, many false statements spread in social networking and bring unnecessary distress to the economic development and social stability. Conversely, the popularity cannot be absolute but needs to take measures.Boost the dissemination of general information at the right node will greatly increase the possibility of poststo become popular information. Therefore, the study to classify and predict the trend of postsis helpful for enterprises and society.

At present, scholars focused on the trend of topics. Scholar Leskovec J[1] discussed the Cascade propagation model of analogy Epidemic Model about topics in blog network. Scholar JingLv[2] considered the degree of active nodesand made propagation model based on the topic of discrete time. K.Y.Chen[3] identified the formation of a popular topic by analyzing the time domain [18]. Some scholars have studiedthe trend of information dissemination. For example, scholar K.Lerman and T.Hogg[4]analyzed user interface and posts' early data to predict the long-term data trends and the popularity of a post. However, scholars rarely involved in the comparative study about the trends of spread betweenpopular information and general information. In this paper, the authors grabbed data from international observation section in Tianya website and then classified the posts in accordance with popularity. At last,we predicted the trends of replies at the set time.

However, the information diffusion models promoted in previous studies do not apply to Tianya website. Therefore, scholar K.Lerman and T.Hogg[4] had used a linear regression model on a logarithmic scale to minimize the residual between actual and predicted values. After that, they estimated the replies at the set timebased on early replies. The authors will analyze all popular posts and ordinary posts by the log-transformed

*Corresponding author.Email: 1973887130 @qq.com(Jingwen Chen) , lsm940425@163.com(Shumeng Liao)

linear regression model. In recent years, the common method to distinguish a popular topic or popular information was mainly based on clustering analysis. Moreover, previous studies often distinguished the popularity of topics or information by seeking the maximum daily number of replies[5]. Scholar Wu[6] believed that the allocation of human's attention is asymmetric. Most people focused on a very few contents, so there will be a small part of posts can attract most attention. In this paper, the division of posts is mainly based on the 20%-80% rule. It is that the former 20% posts which have the maximum number of replies are popular posts and the threshold value is the criterion to measure popularity. In summary, the authors will use the 20%-80% ruleto divide all posts in the international observation section and applythelog-transformed linear regression model to analyze all the posts, not a post, and compare different trends of replies.

This article is organized as follows: in Section 2 we definite popular posts and confirm the log-transformed linear relationship between early reply and dependent variable; Section 3 introduces the linear regression model; in Section 4 we introduce the data capture method briefly and show the results of prediction by figures and tables; Section 5 interprets data analysis results, elaborate theoretical and practical significance; in Section 6 we propose a general conclusion and point out directions for future research.

2. DEFINITION AND INITIAL FORECAST

2.1 Definition of popularposts

At present, scholars had made some researchesto measurepopular topics. Scholar E.Zhou[7]considered some factors in mind such as blog replies, hits, user engagement, etc. to determine a popular topic. ScholarK.Yu[8] thought popular topics were continuity and tended to appear in multiple sections. Based on the above studies, the authors believed that popular posts were able to attract a lot of attention and brought users to interact in a short time. Popularposts were often with a large amount of hits and replies, but the more hits did not represent the popularity. The reason was that some posts were serialized novels with a long time span. So the authors removed these posts when filtering the data. In summary, it was reasonable that maintained replies as the criterion of popularity. Therefore, the authors wouldarrange replies of 48 hours in descending order and as the former percentage of the posts werepopular posts.

Figure 1 showed the frequency of replies based on the current data of posts in 48 hours. X-axis represented replies of the 48th hour, Y-axis represented the frequency.A large number of replies were within 50, while only a few posts had a lot of replies ranged from 200 to 600. The curve near the Y-axis obtainedmany replies, while near the X-axis had a converse result. Some scholars had concluded that community posts showed the long-tailed distribution. It meant thatonly a small part of posts could get a lot of attention, on the contrary, most posts only had a small amount of attention. Therefore, the curvewas consistent with the long tail theory. Specifically, there were about 20% ofall posts with a large number of replies. So it was reasonable and feasible to definite the former 20% percentage of the posts as popular posts and the threshold value was the criteria to judge the popularity.

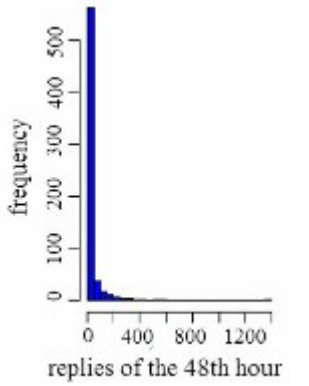


Figure 1.The frequency of the 48th hour replies

2.2Initial forecast

According to the definition of popular posts, this article chose the reply as the sole variableto measure different popularity of posts and applied statistical software SPSS to analyze the original data by linear regression analysis. The purpose was to explore the linear relationship between replies in early time and replies at the 48th hour, but the linear relationship was not obvious. However, previous scholars hadtransformed the original data by logarithmic transformation which showed a significant linear relationship. Therefore, the log-transformed linear relationshiphad been shown in Figure 2. The authors only showedtwofigures which represented the relationship between the replies of the 6th hour (the 12th hour) and the 48th hour. Obviously, the linear growth relationship between the initial replies and the 48th hour replieswas more and more significant with the initial time increased. In summary, the initial log-transformedreplies and the 48th hourlog-transformed replies had a significant linear relationship. Therefore, the authors would choose a log-transformed linear regression analysis model in later article.

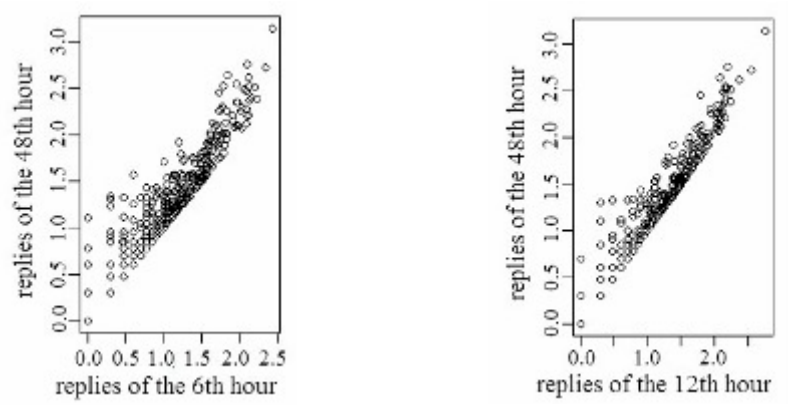


Figure 2. The relationship between 6th hour (12th hour) and 48th hour replies(log-transformed)

3. MODEL

The main research of this article is to explore the overall spread trend of predictedreplies. Scholars had confirmed that using the log-transformed linear regression model to forecast the trend was more accurate. The linear regression model wasbased on the ordinary least squares estimate.

In linear regression model, the dependent variable is $y(t_j)$ which represents the replies of posts at time t_j.The predictive value is $\hat{y}(t_i, t_j)$ and the independent variable is x_i which represents the replies at time t_i. The linear regression model of sample is as follows:

$$y(t_j)=\beta_0+\beta_1 x(t_i)+e$$

The sample regression equation of the linear regression model is

$$\hat{y}(t_i,t_j)=\hat{\beta}_0+\hat{\beta}_1 x(t_i)$$

Where $\hat{\beta}_0$ and$\hat{\beta}_1$are the estimates of regression parameters, $\hat{\beta}_0$is the intercept in the y-axis, $\hat{\beta}_1$is regression coefficient, $e=y_j-\hat{y}(t_i,t_j)$is the residual, which is the vertical distance between dependent variable $y(t_j)$and sample regression line$\hat{y}(t_i,t_j)$.Based on the ordinary least squares estimate (OLSE), the linear model of error is as follows:

$$OLSE = \sum e^2 = \sum [\hat{y}(t_i,t_j)-y(t_j)]^2 \quad (1)$$

Linear regression model used in this paper has been transformed on a logarithmic scale which is based on model (1). Therefore, the linear model of error which has been transformed on a logarithmic scale is as follows:

$$OLSE^* = \sum e_c^2 = \sum [\widehat{\ln y}(t_i,t_j)-\ln y(t_j)]^2 \quad (2)$$

In equation (2),$\widehat{\ln}y(t_i,t_j)=\beta_0^*(t_i)+\beta_1^*\ln y(t_i)$. It is necessary to restore the transformed data to the final predictive value, and here has awidely used model for transformed data. Therefore, the most suitable non-transformed estimated results are as follows:

$$\hat{y}(t_i,t_j)=\exp[\widehat{\ln}y(t_i,t_j)+\delta^2/2] \quad (3)$$

Here$\delta^2=var(e_c)$, the consistent estimate for the variance of the residuals on the logarithmic scale. This article will use the statistical software SPSS to transform data and estimate the variance of the residuals and predictive values. And then apply Eq. (3) to obtain the final predictive on the original scale.

4. DATA ANALYSIS

4.1 Data capture

Tianyawebsiteis the largest Chinese community which has 47 sections totally and users can post messages and replies in any section. After observation, the authors found the sections of Entertainment gossip, Zatan and International observation ranked the top three in Tianya website. The three sections obtained a high degree of attention and popularity. Finally,we selected the international observation section and carried out further research. Here were two reasons, first,the post content and the section title fitted very well; second, this section had a few ads.

The authors used Java language to write a program and implementWeb Crawler to capture data. We required the Web crawler to capture the latest posts in international observation section without disturbing the normal operation of the Internet. After capturing several times, we found that the life cycle of posts in international observation section was generally 48 hours and after 48 hours the growth rate of hits and replies were very small. To ensure the posts had finished the length of life cycle, we set the capture time from 2014.10.23 to 2014.10.30 and the interval was 15 minutes. In the end, the cumulative number of postswas 645.

4.2 Data forecast

This article showed the trend figures of predicted values which contained two parts, popular posts and ordinary posts. The replies of posts were arranged in descendingorder and we selected the former 20% to be popular posts (here were 133 posts whose replies were 32), the remaining 512 posts were ordinary posts.

The authors used statistical software SPSS to analyze the linear relationship between the log-transformed replies of the 1th to the 47th hour and the log-transformed replies of the 48th hour orderly. We obtained the relationship between the log-transformedreplies in early time and the 48th hour replies. Also, we got predicted values and standard deviation of residuals. The following tables showed results for the 6th hour of popular posts.

Table1.The regression coefficients of 6th hour replies(log-transformed)

	Unstandardized coefficients		Standardized coefficients		Frequency
Model	B	Std. Error	Beta	t	Sig.
(Constant)	.755	.222		3.393	.001
Tr-replies(6h)	.953	.058	.822	16.434	.000

a. Dependent Variable:Tr-replies48h

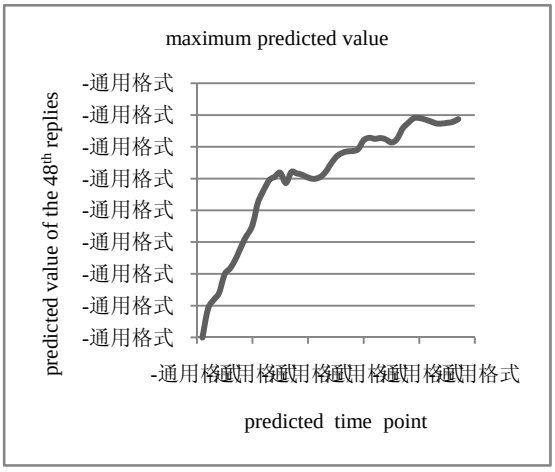
Table2.The residual of 6th hour replies (log-transformed)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.0757	6.1034	4.3522	.65015	132
Std. Predicted Value	-3.502	2.693	.000	1.000	132
Residual	-.59141	1.53524	.00000	.45108	132
Std. Residual	-1.306	3.390	.000	.996	132

a. Dependent Variable: Tr-replies48h

The predicted values of linear regression had included the maximum predicted value, the minimum predicted value and the mean predicted value which had showed in table 2. Therefore, the authors drew the trends of the three predicted values. Fig.3 showed the trend of maximum predicted replies for popular posts. The horizontal axis represented the predicted time point (from the 1th hour to the 47th hour) and the vertical axis represented the predicted value of the 48th replies. In this curve, before the 18th hour, the predicted value showed steep growth trend which was similar to linear trend, and the number was less than 1000; while after the 18th hour, the curve of predicted value grown slowly to 1400, specifically,predicted value between the 18th and 22th hour fluctuated in 1000, between the 22th and 40th hour showed approximately linear growth trend, between the 40th and 47th hour remained stable and unchanged.

However, the trends of mean and minimum predicted value were similar to each other and the replies were both small. Trend of mean predicted value forpopular posts was shown in Fig.4. In this figure, the trend curve was quite steep and reached a peakat the first hour. After the first hour the curve exhibited a slow decline and the number dropped to about 75.



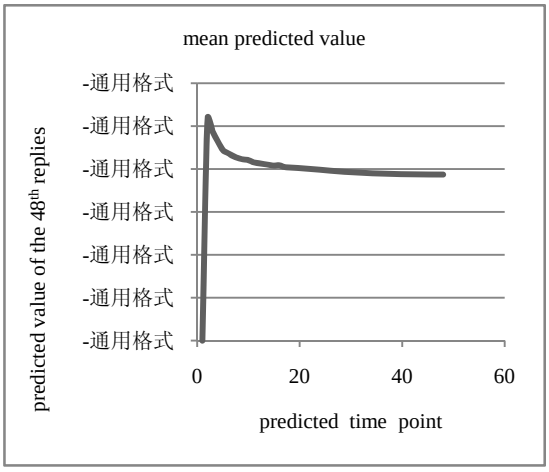
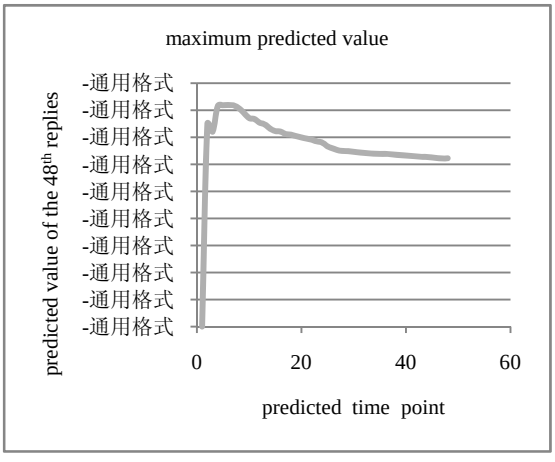


Figure 3. Max-predicted for popular posts **Figure 4. Mean-predicted for popular posts**

Fig.5 showed the trend of maximum predicted values for ordinary posts. Specifically, before the 5 hours, the predicted value showed steep growth trend which was similar to linear trend and the number reached the peak 42; while after 5 hours the curve exhibited a slow decline and predicted value dropped to about 32. Trend of mean predicted value for ordinary posts was shown in Figure 6. In this figure, the trend curve was quite steep and reached a peakat first hour. After the first hour the curve exhibited a slow decline, specifically, after 10 hours the curve was substantially parallel to the horizontal axis and maintained the number 6.



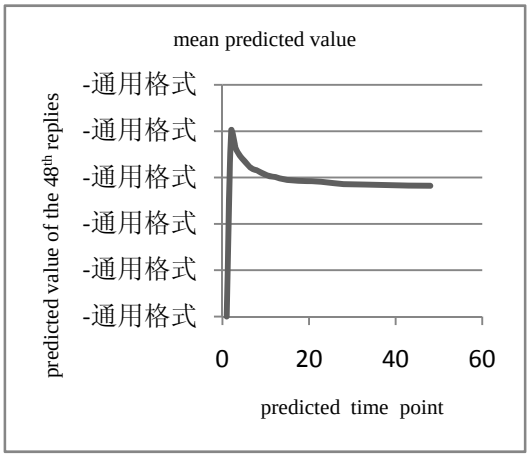


Figure 5. Max-predicted for ordinary posts **Figure 6. Mean-predicted for ordinary posts**

After comparing, the curves of mean predicted value for popular posts and ordinary posts were quite similar, because they both increased steeply at first and then declined slowly. The first hour was an important turning point. After the first hour, the two curves of the mean predicted value showed a downward trend and there were 10-fold difference between the two peaks. When comparing the maximum predicted value, before 5 hours the predicted values of popular posts and ordinary posts showed a sharp rise, after that with the decrease values of ordinary posts and the increase values of popular posts, a huge difference between the two types of posts was significant.

5. DISCUSSION

5.1 Discussion of data analysis

According to the trends of predicted replies for popular posts and ordinary posts, we obtained similarities and differences between the two types of posts. The authors chose the mean and maximum values to discuss, and omitted to describe the minimum predicted replies. The reason was that mean value could represent overall trend of the posts. And the maximum value was significantly reflected the trend of postsbecause of the large difference between threshold value and maximum value.

For popular posts, the trendof mean values rose for a while and then showed a downward state. It can be attributed to two reasons: first, the special mechanisms of Tianya website. In Tianya, when each post increaseda number of replies, this post would automatically appear in the default column. The authors had done a preliminary survey on a small-scale, the results showed that when visiting the forum, users were more inclined to browse the posts in default column. Therefore, it greatly increased the exposure rate of posts when the post gained more replies in the first hour. Based on the linear relationship between log-transformed early replies and the 48^{th} hour replies, when replies in the first hour were large, the predicted value of the 48^{th} hour was consistent with them. However, not all posts would have continued exposure rate as time went on, therefore, some lower exposure postswhose replies grew more slowly or even had no growth, while higher exposure posts still had a lot of replies. The huge differences among these posts leaded to the mean predicted value no longer rise but drop and tend to an average level. Second, select different forecast time points. According to the linear relationship between log-transformed early replies and the 48^{th} hour replies, the closer the early time near to the 48^{th} hour, the more significant the linear relationship would be. Therefore, when forecasting the 48^{th} hour replies at the first hour, the residuals was not the smallest, that was to say the error was much larger. So the predicted results of the first hour may deviate slightly. The error would become gradually smaller as predicted time went on and the predicted results were closer to the actual results. In summary, trend of the mean values shown a downward state after rising for a while.Then it was necessary to explain the 20^{th} hour for popular posts. The 20^{th} hour was an important time point. Based on the existing data, replies of most posts grew more slowly or even had no growth. So after the 20^{th} hour, the curve maintained the state of increasing slowly which was notsame with before.

For ordinary posts, explanation for trend of mean predicted value was same to the popular posts. It is agreeable to explain the trend of the maximum predicted value from the special mechanisms of Tianya website. More possibly, ordinary posts obtained some replies in the first five hours because of the mechanism of default column. In default column, the formerlyranked posts were changing rapidly, while the posts on the behind would be gradually replaced. Therefore, the most possible reason was that usersviewedall posts in default column which was helpful forobtaining some replies in early time, and this behavior really increased the exposure rate. However, as time went on,ordinary posts might be disappeared because of lowerexposure rateandno growth replies. In summary, popular posts and ordinary posts had the difference which was that the maximum predicted value for ordinary posts would show a downward state after rising.

After comparing the difference between popular posts and ordinary posts, we got the conclusion that the

biggest difference between two types of posts was the early replies. Within the first five hours after releasing posts, we could take advantage of the strengths or other ways of the posts to increase replies. These methods may increase the exposure rate in a maximum way and help the early replies greatly improved. Five hours later, the increasing of replies depended on the exposure rate after these posts being pushed to the default column automatically.On the basis of early replies, the exposure rate and replies would continue increase and the possibility of disappearing would greatly reduce.

5.2Theoretical contribution

Current scholars definedpopular posts verily.Some defined the maximum daily replies as the threshold value ofpopularity, and some thought that the user participation would be the standard to measure the popularity. In this paper, we showed the threshold value by figures which exhibited that it was reasonable to use the 20%-80% rule and definite the former 20% of maximum replies as popular posts. Therefore, this method maymake up the gaps in classification field. Based on the mean value of posts, we compared the different predicted values of popular and ordinary posts. In other words, the authors studied the trends of maximum, mean and minimum predicted values at the same time. Obviously, the trends of maximum predicted values for two types of posts had significant differences which provided a theoretical support for the future research.

5.3 Practical contribution

With the rapid development of networks, users could not pay their attention on all contents. Therefore, it is helpful for the enterprises to disseminate information rapidly when they maintain optimistic replies of posts within five hours after releasing. Based on the operation mechanism of Tianya website, when each post increased numbers of replies, this post would automatically appear in the default column.So on the one hand, enterprises can increase the replies in some ways at early time which is helpful for increasing exposure rate; on the other hand, due to the operation mechanism, companies can maintain larger hits and replies after five hours which is based on the previous replies. More importantly, in this way enterprises could save some human and material costs.For forum administrators, these conclusions can provide some help for administrators to screen popularposts by estimating predicted replies. If one post had the similar trend to the popular posts,administrators can recommend it to the popular posts section for usersobtaining the latest and most popular news quickly.

6. CONCLUSION AND FUTURE RESEARCH

6.1 Conclusions

Recalling the whole article, the main content is to study the trends of predicted replies for popular and ordinary posts. The authors capture the latest posts in international observation section and definite the 48^{th} hour replies as the dependent. Here we define the former 20% posts as popular posts whose replies are larger. After comparing the trends of predicted values for popularposts and ordinary posts, we get the conclusion that popular posts increase the replies in some ways within the first five hours after releasing and then dependon users and operation mechanism to increase exposure rate constantly.This paper has two contributions: one is that proposing a new method to distinguish popular posts and ordinary posts. The method is 20%-80% rule. The other contribution is drawing the trends of three predicted value for the two types of posts, and this could provide a reasonable proposal for enterprises and administrators to identify and push popular posts.

6.2 Future Research

In future research, scholars can add some other independents in the linear analysis model, such as hits, published time and so on. More than that, the content of posts canalso be an independent which needs to use the content split method to quantitatethe data in order to study furtherly. It is a good choice to compare different sources of data. So scholars can collect data from Youkuwebsite and analyze the differences between Youkuwebsite and Tianya website.

REFERENCES

[1]Leskovec J, McGlohon M, Faloulsos C, Glance N, Hurt M.(2007). Patterns ofcascadingbehavior in large blog graphs. In: Proc. of the SIAM Int'l Conf. onData Mining. New York: ACM Press, 551-556.

[2] Jing Lv. (2013). Analyze the propagation model for network popular topics. Master Degree Thesis. Beijing: Beijing University of Posts and Telecommunications,

[3] KYChen, LLuesukprasert, SCTChou. Popular topic extraction based on timeline analysis and multidimensional sentence modeling. (2007).IEEE Transaction on Knowledge and Data Engineering, 19(8):1016-1025.

[4] K Lerman, T Hogg. (2010).Using a model of social dynamics to predict popularity of news.www.arxiv.org,

[5] FeiXiong. (2013). Analyze the internet users' behavior and information evolution model. Master Degree Thesis. Beijing: Beijing Jiaotong University.

[6] FWu, BAHuberman. (2007). Novelty and collective attention.Proceedings of the National Academy of Sciences of the United States of America, 104(45).

[7] E Z Zhou, N Zhong, Y F Li. (2011). Popular topic detection in professional blogs. Active media technology, 6890:141-152.

[8] K Y Chen, L Luesukprasert, S C T Chou. (2007). Hot topic extraction based on timeline analysis and multidimensional sentence modeling. IEEE Transaction on Knowledge and Data Engineering,19(8):1016-1025.

[9]ZFZhang,QDLi. Popular topic discovery and trend analysis in community question answering systems. (2011).Expert Systems with Applications, 38(6):6848-6855.

[10] JFYu, YQHu, MYu. (2009).Analyzing netizens'view and reply behaviors on the forum. Physica A, 389-398.

[11] XMSi, YLiu.Empirical analysis of interpersonal interact behavior in virtual community. (2011). ActaPhysicaSinica,60(7):078903.

[12]GSzabo, BAHuberman. (2008). Predicting the popularity of online content. www.arxiv.org,

[13]KLerman,AGalstyan. (2008). Analysis of social voting patterns on Digg. www.arxiv.org,

[14]FDing, YLiu, HCheng.(2010). Read and reply behaviors in a BBS social network.Proceedings and International Conference on Advanced Computer Control.Shengyang.

[15] S Jamali, H Rangwala. (2009). Digging Digg: Comment mining, popularity prediction and social network analysis. International Conference on Web Information Systems and Mining,Shanghai, 32-38.

[16] X M Si, YLiu. (2011). Empirical analysis of interpersonal interacting behavior in virtual community.ActaPhysicaSinica,60(7):078903.