Research on Hot Issues and Evolutionary Trends in Network Forums

Li-Jie Cui[1], Hui He[2] and Wei Liu[3]

[1]*School of Software and Engineering, Harbin University of Science and Technology, Harbin, China*
[2]*Harbin Institute of Technology, Harbin, China*
[3]*Harbin Institute of Technology Software Engineering Co. Ltd, Harbin, China*
andyclj1977@163.com, hehui@hit.edu.cn,dicklw@163.com

Abstract

The network forum is an important channel of information dissemination for Chinese Internet users and has become the primary source of opinions concerning "hot issues." This paper presents a fast and efficient method to mine "hot posts" in network forums and to analyze the evolutionary trends of hot issues. The proposed method uses statistical and mathematical models to sort posts into sections. The proposed method is also used to analyze the evolutionary trends of hot posts, thus determining whether tracking is necessary for a certain post. Experiments were conducted to verify the feasibility and effectiveness of the proposed model. This study establishes a foundation for the further study of the evolution of hot issues in complex networks.

Keywords: *network forum; hot issue; public opinion; information dissemination*

1. Introduction

In the current information era, the Internet has become an essential way to obtain news, knowledge, and other information for work, education, or entertainment. Among the many channels of information dissemination on the Internet, the network forum has become an important medium of transmitting information for Chinese Internet users. Taking the Tianya Club as an example, tens of thousands of Internet users post topics in which they are interested in. Chinese users express their opinions regarding every day topics in this forum. Thus, the influence of the network forum is broad and far-reaching. "XiaoYueYue" is a post released by an Internet user in Tianya Club in May 2009. Merely a few days after the release of the post, hundreds of thousands have viewed it and tens of thousands have commented on it. The name "XiaoYueYue" has become popular in China and even become a mantra for some university students. Similarly, the post "my dad is Li Gang" in 2010 has become a popular household incident, and many users have debated on this topic for days in the network forum. Different views have been posted regarding the event, and people have participated in the intense discussion. These posts are extraordinary and have circulated in the Internet for a long time; until now, these posts have not yet been completely forgotten. In addition, if a new policy is quickly implemented by the government, then the pros and cons of this policy are considered and analyzed by a large number of Internet users in related sections of network forums. Thus, this policy becomes better understood by the citizens in a short span of time. The network forum plays an important role in the dissemination of information. However, not everything is absolute. We should analyze this channel of information dissemination from a dialectical perspective. The emergence of information technology has led to numerous problems and even to some disagreements. The key to solving these problems is determining the optimal solutions for these problems and figuring out the best

way to deal with contradictions. Freedom of speech must still be maintained in network forums; however, this freedom of expression must not violate the law. Actions violating public morality or laws are criticized in network forums, which prompt people to react and formulate their opinions. Public opinion helps in the timely correction of negative behavior and ensures that justice is applied equally. Treating network forums as venues for spreading negative remarks and illegal speech greatly affects or even cause serious harm to the community. Network forums should abide the law. Consequently, Internet netizens, which are the main components of network forums, should consciously maintain the harmonious atmosphere [1] in the network forums. Illegal behavior in network forums can be rectified by analyzing and controlling public opinions through technology.

The core elements of information dissemination and public opinions in the Internet include hot issues, focus, sensitive points, frequency, *etc.* Here, "hot issues" refer to topics subjected to moral judgment [2]. Hot issues are information points that are given attention over a period of time. The topic "XiaoYueYue" mentioned above is a hot issue. Hot issues are problems that trigger the concern and interest of people. Moreover, hot issues greatly affect the community and usually reflect the concerns of the members of the society. Hot issues result in high-risk social operations because of the various consequences these issues generate. These issues even affect social stability and may impede the construction of a harmonious society [3]. Thus, to control public opinions, we should track, search, and regulate hot issues to safeguard social fairness and justice [4]. This research on the development and evolution of hot issues in network forums has academic significance and social value [5].

2. Principle of related methods

2.1. Data collection

The Tianya Club has been respected by the global Chinese netizens since its establishment in March 1999 because this online community is open to all, inclusive, and contains a wide range of issues and concerns [6]. After decades of development, Tianya Club has become a communication platform that combines online forums, social media, and blogging.[7] This online community also includes a series of functional services, such as personal space, albums, music box, classified information, station message, virtual store, "come on" bar, question and answer, corporate brand home, *etc*., [8]. Tianya Club has become a comprehensive virtual community and a large social networking platform and has a significant influence on the public opinion in China. Thus, taking Tianya Club as an example, we propose a fast and effective method of mining hot posts in Tianya Club and tracking the evolutionary trend of these posts [9].

We have mined a total of 18753 posts from February 2010 to December 2010 in the "People Voice" section of the Tianya Club [10]. Each post includes the following information: post title, subordinated section, number of visits, number of replies, author, publication date, post contents, and comment replies. Among these data, we focus on the number of visits and number of replies of posts [11].

2.2. Data analysis

We have extracted the number of visits and number of replies for each post whose values constituted a tuple (number of visits, number of replies). We can obtain information regarding the *i*th post by checking the tuple. The extracted attributes, namely, number of visits and number of replies, are projected onto the following Cartesian coordinate system (Figure 1):

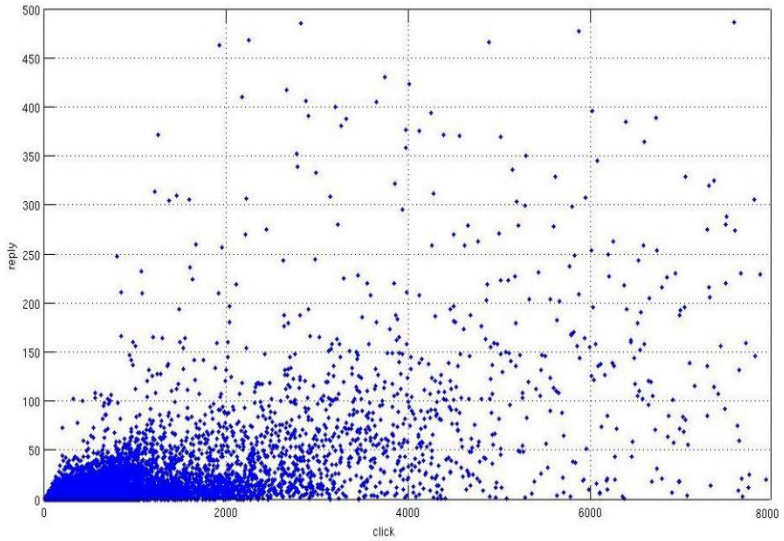


Figure 1. Distribution of number of visits versus number of replies

In Figure 1, the *x*-axis represents the number of visits, whereas the *y*-axis represents the number of replies. An analysis of the graph shows that majority of the number of visits and number of replies are within the interval [*x*: 0, 200; *y*: 0, 50] of the coordinate system. An increase in the number of visits and number of replies in the upper right quadrant corresponds to a decrease in number of posts in the same quadrant. The number of hot posts should be lower than the total number of posts. Thus, posts close to the upper right quadrant are popular. In general, the distribution of number of visits and number of replies adhere to the following patterns:

1. Few visits. In this case, the topic has failed to catch the interest of website visitors, thus resulting in a small number of visits. These types of posts are called "cold posts," which are regarded as meaningless posts.

2. Large number of visits, but few replies. This case corresponds to the lower right quadrant of Figure 1. In this situation, we cannot evaluate whether a post is a hot post by using only the number of visits. Posts in this region have large number of visits, but very few replies. The topics of these posts may have caught the attention of a large number of website visitors who are interested only in browsing, but not in participating with the topic discussion. These types of posts are called "pseudo-hot posts."

3. Large number of visits and replies. These posts are located in the upper right quadrant of Figure 1. These posts are the "hot posts" needed in this study. Hot posts have large numbers of replies and are popular for a long duration. Thus, these posts have caught the attention of a large number of website visitors who are also interested in participating with online discussions. People express their views and opinions on a topic, which may lead to the discussion of other social problems. Therefore, many negative or unlawful remarks are generated. Considering the fast dissemination of information through the Internet, failing to censor these unlawful remarks may negatively affect the society. Thus, hot posts should be properly regulated to avoid negative consequences. However, regulations concerning public opinion should be appropriate, rational, and acceptable.

By analyzing these three patterns, we have constructed a framework that classifies crawled posts and then extracts the hot posts. We have also created a list of standard qualifications that are required of hot posts. An in-depth analysis of hot posts is conducted below.

2.3. Methods and results

We assume that the number of visits and number of replies are two random variables and that more number of visits results in more number of replies. For simplicity, we assume that posts with a small number of visits cannot have a large number of replies. The Pearson product–moment coefficients of the two attributes are calculated in 18 753 posts. By using the calculation results, we can determine that $r_{x,y} = 0.5$. This value indicates that the two variables exhibit a high correlation between each other. The ratio of the number of visits to the number of replies is also an important factor. This ratio reflects whether debates may ensue during the discussion of a topic. A post with a high ratio of number of visits to number of replies denotes that increased effort should be given when tracking and controlling the post, provided that this post contains a high number of visits; otherwise, the calculated ratio is insignificant. For example, analyzing posts with one visit and one reply will result in insignificant data.

Our ultimate goal is to extract "hot posts" that we deem valuable among all posts in a given section. We can also extract other interesting information, such as whether the authors of these hot posts play the role of opinion leaders and whether their published posts have strong influence on other people[12] .To obtain these data, we can divide the people involved in these posts into several camps and observe the camp evolutionary trend. However, we have not included these topics because of space limitation. To extract hot posts, we can score all the posts and then select the top *N* posts according to need. The following scoring formula is used in scoring these posts:

$$S(p_i) = \omega_1 \frac{x_i}{average(x)} + \omega_2 \frac{y_i}{average(y)} + \omega_3 \frac{y_i / x_i}{\max(\alpha)} \tag{1}$$

where $S(p_i)$ represents the score of the *i*th post, $average(x)$ represents the average number of visits, $average(y)$ represents the average number of replies, $\max(\alpha)$ represents the maximum ratio of the number of visits and number of replies in all the tuples, and $\omega_1, \omega_2, \omega_3$ represent the weighting factors. Considering that the number of replies can better reflect the probability whether a post may result in a debate, we assume that $\omega_1 < \omega_2$. We can consider a third factor if the scores of the posts cannot be determined by the number of visits and number of replies.

By adjusting the proportion of the number of replies, we can obtain:

$$0 < \frac{y_i / x_i}{\max(\alpha)} < 1 \tag{2}$$

Here, ω_3 is larger than the other two weighting factors. After testing the "People Voice" section, the following initial settings can be established to obtain satisfactory results: $\omega_1 = 0.1$, $\omega_2 = 0.2$, $\omega_3 = 0.8$. The weighting factors are set differently and trained through machine-learning methods because the topics discussed in different sections have varying characteristics.

The following results are obtained as Table 1:

Table 1. The test results of different topics

Number of visits	Number of replies	Score	Title
25608	2640	28.269312	I deserve it?! More great sorrow of the Chinese society!!
67758	2626	30.813650	How is the secretary grabbing forty million fortune?
34799	3134	33.647907	The Changping court man-made, malicious, protracted
88979	3574	51.665966	I pick up the legal weapons in order to safeguard my legitimate rights and interests.
27874	5030	68.396366	Hangzhou appeared Wu Ying fraud of "Dongyang Edition"!
29062	7894	81.246101	Wife unauthorized use of the children's education fund, angry and want a divorce!

3. Development trend analysis

3.1. Preprocessed data

We then analyze the first post "Unauthorized using of the children's education fund by wife, angry and wanted to divorce!". This article is published with the nickname "Why spend the money" by an Internet user in August 17, 2010. In October 14, 2010, we have observed a total of 29 062 visits and 7894 replies for this topic. Analyzing the trend of this post is necessary to determine whether it is necessary to continue tracking it. The below methods are used.

The printing area is 122 mm × 193 mm. The text should be justified to occupy the full line width, so that the right margin is not ragged, with words hyphenated as appropriate. Please fill pages so that the length of the text is no less than 180 mm, if possible.

We analyze the single post text and extract the publication date and the time of the last reply, in which subtracting the former from the latter yields to the duration of the post. The time of the last reply should be within a certain time interval from the crawling time. If the time of the last reply is close to the crawling time, then we cannot conduct an accurate analysis of the post. Our calculations show that the difference of the publication date and the time of the last reply equals to 58; thus, this thread lasted for 58 days. Groove marks are created for each day by using an array to segregate comments according to publication date. Subsequently, the subscript of the array is mapped to the *x*-axis, and the value of the array

becomes the value of the *y*-axis. Figure 2 shows the discrete points plotted in the Cartesian coordinate system.

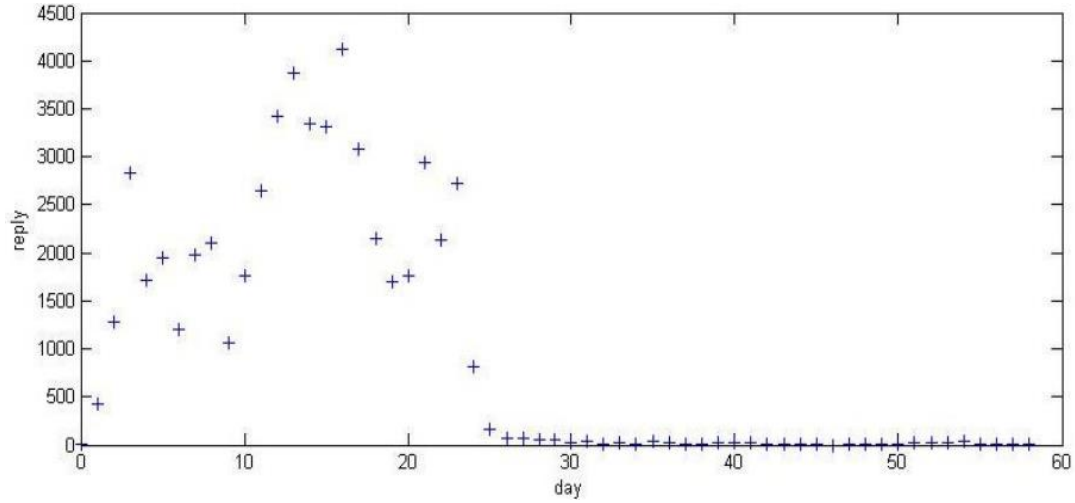


Figure 2. Number of replies per day over time

In Figure 2, the *x*-axis represents the days that passed since the creation of the post, whereas the *y*-axis represents the number of corresponding responses. The popularity of this post tends to decrease and level off, and soon the post "sinks." After removing noise points, we can utilize a suitable curve that fits these discrete points. The result is presented below (Figure 3).

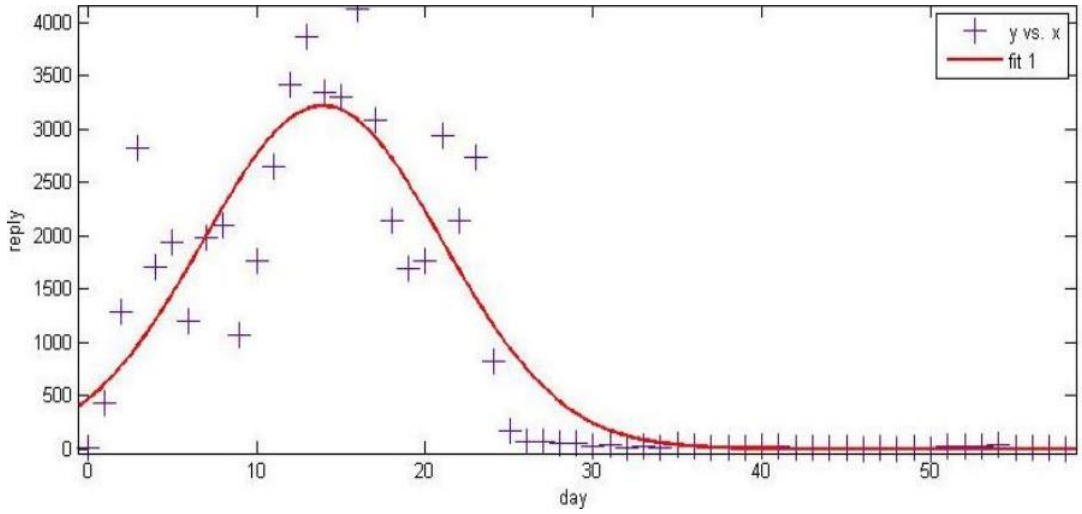


Figure 3. Number of replies per day distributed over a Gaussian Curve

3.2. Selection and evaluation of mathematical models

After testing various curves, we have determined that the Gaussian curve model has the best fit. Thus, we can use the Gaussian curve as a tool for fitting the data curve of other posts.

The derivative of the Gaussian function exists, which denotes that we can use the above scheme. The Gaussian model is expressed as follows:

$$y(x) = a \times e^{-(\frac{x-b}{c})^2} \tag{3}$$

The example is fitted by using the Gaussian model. The following parameter values are obtained:

$$a = 3218(2849,3586)$$

$$b = 13.92(12.99,14.85)$$

$$c = 9.984(8.167,11.35) \tag{4}$$

The above values have a confidence interval of 0.95. By substituting these values into the Gaussian model, we can obtain the following:

$$y(x) = 3218 \times e^{-(\frac{x-13.92}{9.984})^2} \tag{5}$$

The evaluation of the fitting effect is detailed below:

1. $SSE = 1.548e+07$. SSE, which is the sum of squares of an error term, reflects the discrete status of the observed values of each sample. This variable also denotes the sum of the squares of a group or the residual sum of squares.

2. $R-squre = 0.8339$. $R-squre$ is the fitting coefficient. The greater the value of this variable is, the better the fit becomes.

3. $RMSE = 525.7$. $RMSE$ is the root mean square error, which is a numerical index that can be used to measure accuracy.

After analyzing these indicators, we have concluded the fit to be ideal.

3.3. Core analysis method

By deriving the independent variable x from $y = f(x)$, we can obtain $y'(x)$. Let $y'(x) = 0$; then, we can obtain the largest extreme point and set it as x_m. The following two conditions are necessary in analyzing posts:

1. $y = f(x)$ has no extreme points;

2. If $\exists x > x_m$, then $y'(x) > 0$.

If condition 1 is established, then $y'(x) > 0$ or $y'(x) < 0$ and $y(x)$ is monotonic. If condition 2 is satisfied, then $y'(x) > 0$, and $y(x)$ is monotonically increasing. When both conditions are satisfied, the heating degree of the post has an upward trend, and we should continue to focus on the trend and evolution of this post.

If point x_n exists such that $y'(x_n) = 0$ holds true, then x_n is a turning point in the heating trend. We then determine the largest $x_{\max}$ that can establish $y'(x) = 0$. By studying x_t, which satisfies $x_t > x_{\max}$, we can denote that if $y'(x_t) < 0$, then the heat of the post experiences a downward trend. Correspondingly, if $y'(x_t) > 0$, then the heat degree has an upward trend, and we should be concerned on the future trend of this post.

In this case, we set $y'(x)=0$ and obtain $x \approx 14$. Thus, we use $x_t = 16 > 14$ and obtain $y'(x_t) < 0$. We can assert that in the current case, the heat of this post is decreasing.

4. Conclusion

Given its rapid development, the Internet has become the main place for information dissemination. To improve the management and monitoring of the Internet, information and public opinions on the Internet should be collected and analyzed. The emergence of network forums has significantly changed the Web habits of people. These forums have quickly become a place for information dissemination, thereby enabling people to express their views and exchange ideas. Thus, analyzing and regulating online forums could strengthen the management and monitoring of public opinion. Posts are the metacells of network forums and the source of information dissemination. An efficient tracking and regulation method for hot posts is needed to control public opinions.

This paper presents a fast and effective method that can discover hot posts and a mathematical method that can analyze the evolutionary trend of these posts. This paper is designed to lay a foundation and pave the way for the next regulation system. On the basis of the analysis results, the proposed method is rapid, feasible, and can obtain ideal experimental results.

Acknowledgements

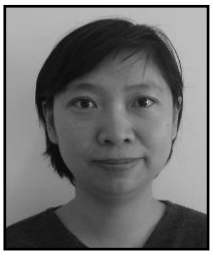
This work was supported by the national high technology research and development plan (2010AA012504、2011AA010705) and the national key basic research and development plan (G2011CB302605、2007CB311101) and the National Natural Science Fund(61173145).

References

[1] L. -H. Wang, "On the network public opinion and the conversion of public opinion and its impact", Tianjin Social Sciences, vol. 4, **(2008)**.
[2] H. -G. Xie and Z. -R. Chen, "Internet content and public opinion depth analysis mode", China Youth Journal for Political Sciences, vol. 3, **(2006)**.
[3] B. Lu, "Automatic discovery and analysis of the Internet public opinion hotspot Research and Implementation", Master's degree thesis of Department of Computer Application Technology of Beijing University.
[4] T. Wenjun and H. Bin, "Stronger Formal Security Model of Three-party Authentication and Key Distribution Protocol for 802.11i", IJSIA, vol. 6, no. 4, **(2012)** October, pp. 163-174.
[5] R. D. Caytiles, D.-H. Wang, D. -G. Oh and B. J. Park, "An Enhanced Packet Buffering Transmission (EPBT) Architecture Design for Performance Enhancing Proxies (PEPs)", IJSEIA, vol. 6, no. 4, **(2012)** October, pp. 155-164.
[6] N. A. Mohota and S. L. Badjate, "Efficient Design of Routing Node to Evaluate the Performance of Network Based Communication Infrastructure for SOC Design", IJHIT, vol. 5, no. 4, **(2012)** October, pp. 61-78.
[7] Z. Na, Y. Cui, Y. Xu and L. Chen, "A Novel Mobility Prediction Algorithm Based on LSVR for Heterogeneous Wireless Networks", IJFGCN, vol. 5, no. 3, **(2012)** September, pp. 93-104.
[8] D. Gupta and A. K. Sharma, "Investigations on Energy Efficiency for WSN Routing Protocols for Realistic Radio Models", IJFGCN, vol. 4, no. 3, **(2011)** September, pp. 61-72.
[9] J. Park, J. Choi, M. Park, S. Hong and H. Kim, "A Study on Intelligent Video Security Surveillance System with Active Tracking Technology in Multiple Objects Environment", IJSIA, vol. 6, no. 2, **(2012)** April, pp. 211-216.
[10] Y. Shi, L. Zhang and L. Zhu, "An Approach to Nearest Neighboring Search for Multi-dimensional Data", IJFGCN, vol. 4, no. 1, **(2011)** March, pp. 23-38.
[11] B. Yun, "ChangBusiness Process Management of Telecommunication Companies: Fulfillment and Operations Support and Readiness Cases", IJFGCN, vol. 4, no. 3, **(2011)** September, pp. 73-86.

[12] D. Wang and W. -Z. Zhang, "Discovery and Research about Forum influences based on the high weight set of words", Micro Computer Information, **(2011)**.

Authors



Li-Jie Cui is a lecture in school of software, Harbin University of Science and Technology, China. She was born on February 1977. She achieved her Master degree in software engineering in 2005 at Harbin Institute of Technology. Her research emphasizes on information technology, software engineering and network security.

Hui He received the B.S., M.S. and Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China. Since September 1999, she has been with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, where she became an Associate Professor in October 2007. Her research interests include network computing, network security.

Wei Liu is a senior engineer in Harbin Institute of Technology Software Engineering Co. Ltd, Harbin China. He was born on October 1976. He achieved his Master degree in software engineering in 2004 at Harbin Institute of Technology. His research emphasizes on information technology, software engineering and network information and security.