In-depth Analysis of Online Hot Discussion about TCM

Yongliang Zhao, Xijin Tang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Beijing 100190, China
{ylzhao89, xjtang}@amss.ac.cn

Keywords: Online Discussion; Behavior Analysis; Power Law; Sentiment & Attitude

Abstract. Online communities generate explosive volume of texts makes online discussions become an interesting research field. This paper takes one hot discussion on traditional Chinese medicine (TCM) crawled from Tianya Forum as an example and makes analysis from two aspects: behavior and opinion. It is found that the time interval between replies and the number of participant's replies follow the power law distribution with the exponent more than 2.0. Online hot discussion has the characteristics of long survival period, many participants and causal language, etc. Then the sentiment analysis method is taken to analyze the debate and the result indicates that the proportion of either positive replies or negative replies is around 45%. We manually label the attitudes of participants whose replies are more than 10. It shows that there is a polarization in the discussion from the view of either sentiment or attitude. Comparing the sentiment with the attitude, we find that each participant's attitude is always clear, while his sentiment is not stable; the sentiment of one reply is always definite, but the attitude of one reply varies. So we need to treat sentiment and attitude differently, for which most researches may not care much.

1. Introduction

With the emergence of Web 2.0 and the rapid development of social media, people express their opinions on societal hotspots and the livelihood issues, etc. more freely and easily. Social media, such as BBS and microblogs provide a platform to the public to discuss issues deeply and broadly. Some issues are no longer limited to expert-level closed meetings, but involved with more communities. It is worth paying attention to those online open discussions.

There have been a variety of researches about group discussions. Tang (2010) proposed two technologies, i.e.: iView and Cormap, to analyze the process of discussion [1]. These two technologies have had a wide range of applications. For example, Tang (2009) used them to analyze public opinions by questionnaires [2]. Luo and Tang (2013) applied iView analysis to study social network and relationship management in mainland China based on papers published on national flagship conferences [3].

Online discussions have some different characteristics compared with traditional group discussions. For example, a thread of BBS is an open discussion. The topic of this discussion is given by the first post of the thread. The participants refer to those who click the thread, read or even reply the contents of the thread. Since only those who reply are recorded, we analyze general behaviors, active contributors, i.e. those who comment to the discussion, etc. In terms of behavior, Celli, *et al*. (2010) provided a quantitative analysis of online behavior in Friendfeed [4]. Guo *et al*.(2012) researched the characteristics of users' behavior in Weibo[5]. Guan *et al*. (2013) carried an empirical study on hot social events on SinaWeibo [6]. Cui, He and Liu (2013) discussed hot issues and evolutionary trends in online forums [7]. Zhao and Tang (2013) studied the pattern of users' behavior in Tianya Forum [8]. In terms of content, much work is about sentiment analysis and opinion mining. Pang and Lee (2008) wrote a review [9]. Some Chinese sentiment ontology libraries are constructed [10, 11]. Wang *et al*. (2013) introduced a sentiment analysis method of different granularity, from terms, sentences to documents[12]. Shi, Wang and He (2013) carried a case study of "7.23 Wenzhou Train Collision" using sentiment analysis of microblogs based on sentiment ontology [13].

In our research, we engage in analyzing those online hot discussions from both behavior and opinion as shown in Figure 1 and focus on online discussion mining over textual data crawled from Tianya Forum. The Forum provides a suitable data source to study the societal issues about daily lives, social unfair, corruption, phenomena of society, etc. For better illustration, we select threads on typical issues of distinct attitudes, one of which is on TCM (traditional Chinese medicine) issue.

The rest of this paper is organized as follows: Section 2 introduces the data processing, carries a detailed study from daily replies, time interval and participants in the discussion about TCM and summarizes the characteristics of online hot discussion. Section 3 studies the opinion in the discussion, including sentiment analysis method and results, the main participants' attitudes towards TCM, and a comparison between sentiment and attitude. Section 4 presents concluding remarks.

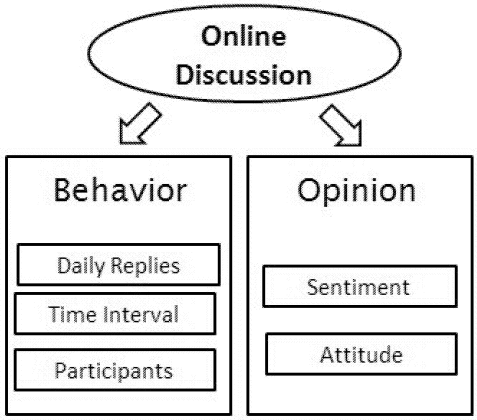

Fig. 1. The analysis of online hot discussions.

2. Behavior Analysis

This section introduces the data processing and carries on in-depth analysis of the discussion about TCM from the aspect of behavior. Daily replies, time interval and participants in the discussion, etc. are studied together with a brief summary about the characteristics of online hot discussion.

2.1. Data Processing

Tianya Forum, which has more than 80 million registered users and is visited by more than 200 million users every month, is a comprehensive online virtual community [14]. It includes many online discussions which cover diverse topics.

A crawler has been developed to download threads from Tianya Forum every day since October of 2010[15]. Among all the threads crawled, there are some hot threads with their clicks more than 100,000 and replies more than 10,000. Obviously those highlighted threads are hot online discussions, which amount to 100 among the crawled data.

We select several threads about TCM as listed in Table 1. The first thread labeled as 2822432 is the hottest one with replies far more than the other two threads [8]. We conduct in-depth analysis to this thread. The thread, with first post published on October 16, 2012, was closed on November 29, 2013 and persisted 410 days. There are 88,698 replies and 4,890 participants involved in the discussion.

Table 1. Hot threads about TCM (accessed on May 1st, 2014)

ID	*Title*	*Posted Date*	*Clicks*	*Replies*	*Participants*	*State*
2822432	Abolishing TCM is good for health.	Oct.16, 2012	491,611	88,698	4,890	Closed
2121178	As the descendants of TCM culture, we cannot help but say something, seeing the posts of abolishing TCM.	Mar.21, 2011	594,867	34,376	5,152	Open
2317943	TCM must exit from the national health system.	Nov.12, 2011	309,312	31,617	5,896	Open

2.2. Daily Replies

The daily replies of the thread are calculated with results as shown in Figure 2. This thread has a total of 88,698 replies with the average daily replies being 216. The daily replies get to a minimum of 0 and are up to 838, which took place on May 4, 2013. The daily replies of the thread vary sharply and are different from that of most hot threads, which usually have a lot of replies during the first few days from the posted date and scattered

replies at the following days. We go further to observe the distribution of daily replies and find that they are subjected to uniform distribution.

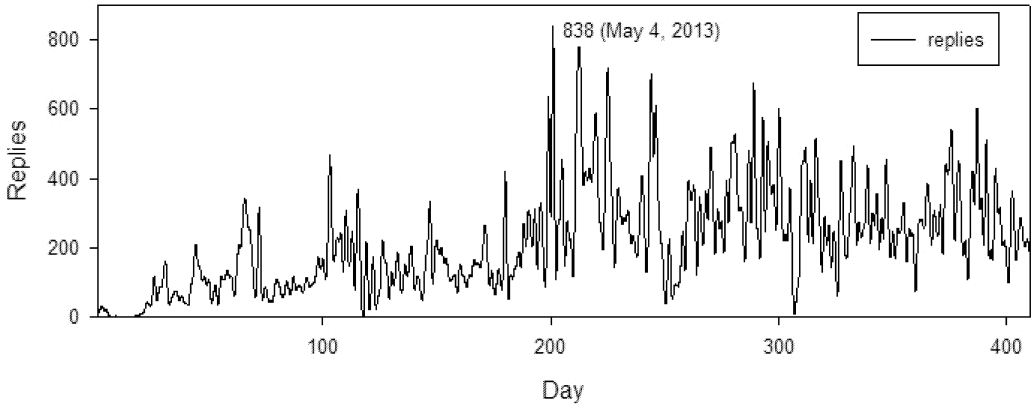

Fig. 2. Daily replies of one hot thread about TCM.

2.3. Time Interval between Replies

The time interval between any two continuous replies is measured as follows: if the time interval is less than 1 hour, labeled as 1; if the time interval is greater than or equal to 1 hour but less than 2 hours, labeled as 2, etc. The distribution of time interval between replies is as shown in Figure 3 with a log-log plot. It is found that the Time Interval 1 appears for the most 87,428 times. And the time interval follows the power law distribution with the exponent being 2.32. The finding supports that human behaviors, especially online behaviors, have the characteristic of eruption in a short period of time [16].

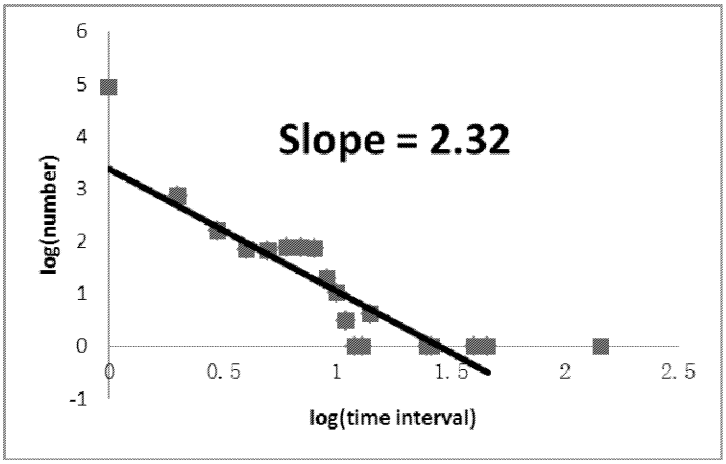

Fig. 3. The distribution of time interval between replies.

2.4. Participants of the TCM Thread

Analysis also goes to participants involved in the discussion. A total of 4,890 participants took part in the discussion and the distribution of participant's replies is as shown in Figure 4 with a log-log plot. The distribution of participant's replies is the power law distribution with the exponent being 2.26, which has the characteristic of heavy tail. The participants, who reply only once, amount to 3,448 and make up 70.5% of all the participants while their replies only account for 3.89% of all the replies. However, the participants, who ranked top 10 by replies with a total of 57,075 replies, only make up 0.2% of all the participants while their replies get to 64.35% of all the replies. It shows that the hot discussion about TCM is dominated by only a few participants, while the majorities are just on-lookers.

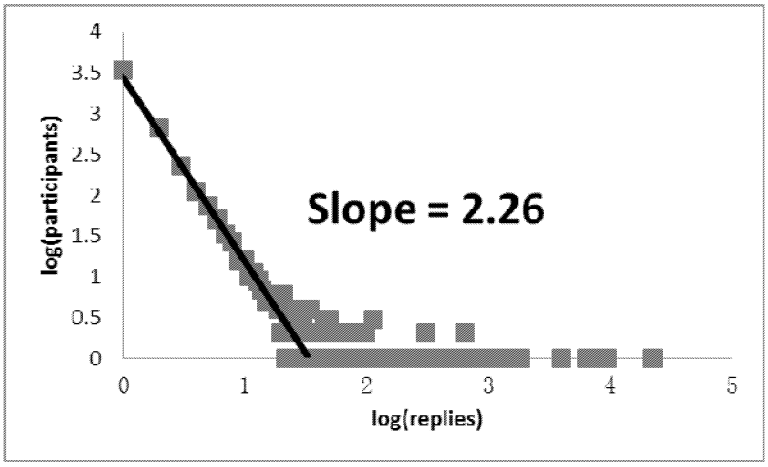

Fig. 4. The distribution of participant's replies in the TCM thread.

The amounts of daily participants and new participants are as shown in Figure 5. It shows that the variation of the numbers of daily participants and new participants is alike. The correlation coefficient between them is 0.89, while the correlation coefficient between the daily participants and daily replies is 0.75. It demonstrates that new participants comprise the majority of daily participants. The more the new participants, the more the daily replies.

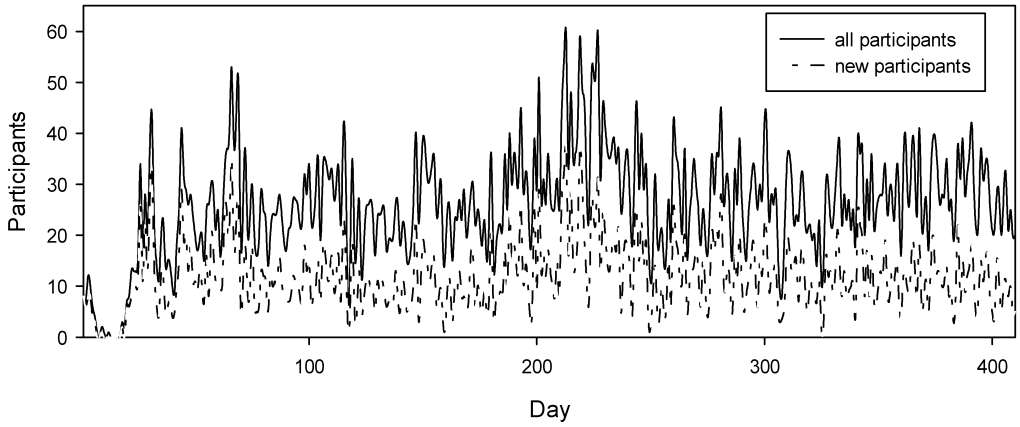

Fig. 5. The amounts of daily participants and new participants in the TCM thread.

2.5. The Characteristics of Online Hot Discussion

Online discussion, which mainly refers to the discussion in BBS, microblogs and other social media, has both advantages and disadvantages compared with traditional group discussions. The easy access and opening to the public mean that online discussion be joined by anyone at any time and any places, which enables online hot discussion to last long, have many participants and produce great influence. For example, the thread about TCM lasts 410 days with a total of 4,890 participants, which hardly happens in one normal discussion, such as meeting, seminar, etc. As to disadvantages, participants with different backgrounds, knowledge, etc. take part in the same discussion may lead to irrational arguments and accusations.

3. Analysis on Sentiments and Attitudes

Besides general behaviors, we go further to conduct sentiment analysis and check participants' attitudes toward TCM in the discussion.

3.1. Sentiment Analysis

We first address the process of sentiment analysis in Chinese context and then take the relevant analysis of the discussion about TCM.

Procedures. The sentiment analysis procedure is based on sentiment ontology, and conduct analysis at four levels based on computation unit, the level of documents, the level of sentences, the level of phrases and the level of terms. Here we use the sentiment ontology libraries constructed by Xu *et al*. (2008) as negative sentiment ontology and positive sentiment ontology [10]. And the privative-terms ontology and degree-terms ontology are provided by Hownet [11]. The sentiment of each reply in the thread is calculated by the following steps:

Step 1: Divide the reply (document) into sentences and each sentence into terms by Chinese word segmentation tools. For each term w in one sentence, repeat:

If $w \in$ sentiment ontology, then $w \in$ sentiment terms of that sentence;

else compute the similarity SW between w and the basic terms [17]. If $SW > threshold$, then $w \in$ sentiment terms.

Step 2: For each sentiment term $w_i (i = 1,.., N_w)$ in one sentence, if the sentiment term is decorated by degree-terms w_a or privative term w_b, then the sentiment value O_{w_i} of the term w_i is computed according to Equation (1):

$$O_{w_i} = M_{w_a} * M_{w_b} * O^s_{w_i} \tag{1}$$

where M_{w_a} is the value of degree-terms, M_{w_b} is the value of privative term, and $O^s_{w_i}$ is the original value of the sentiment term w_i.

Step 3: For each sentence $s_j (j = 1,..., N_s)$ in the reply, compute its sentiment value O_{s_j} by summation of the values of all sentiment terms in the sentence by Equation (2):

$$O_{s_j} = \sum_{i=1}^{N_w} O_{w_i} \tag{2}$$

Step 4: Compute the sentiment value of the reply by summation of the values of all sentences by Equation (3):

$$O_d = \sum_{j=1}^{N_s} O_{s_j} \tag{3}$$

Step 5: Judge the sentiment polarity of the reply by Equation (4):

$$O_d \begin{cases} > 0 : positive \\ = 0 : neutral \\ < 0 : negative \end{cases} \tag{4}$$

Above procedure is one of the most used methods for normal texts [13, 18, 19], unlike short texts in the microblogs where sentiments are simply calculated by sentiment terms matching.

Results. The value of each reply in the discussion about TCM is calculated using the above algorithm and the result is as shown in Figure 6. In the total of 88,698 replies, positive replies reach 38,650, accounting for 44% of all the replies; while negative replies get to 40,167, making up 45% of all the replies. In this case, positive replies and negative replies keep roughly the same; the whole discussion displays 2-polarization, which often happens among public debates [20].

The ratio of daily positive replies and negative replies are as shown in Figure 7. In daily replies, the quantities of positive replies, neutral replies and negative replies vary along with the total amount of replies. However, the proportion of neutral replies keeps roughly the same every day, maintaining at around 10%, while the other two cover around 90%. Positive replies and negative replies vary along the opposite direction beside the value of 0.45.When the ratio of positive replies is high, the ratio of negative replies is low, and vice versa. This shows that the sentiment of daily replies mainly depends on the relative proportion of positive and negative replies.

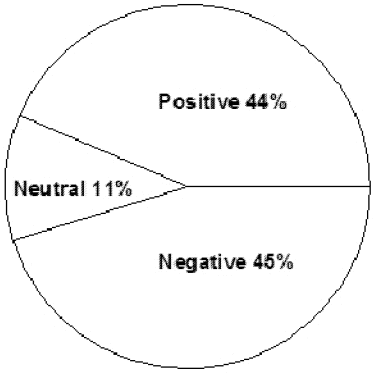

Fig. 6. The result of sentiment analysis.

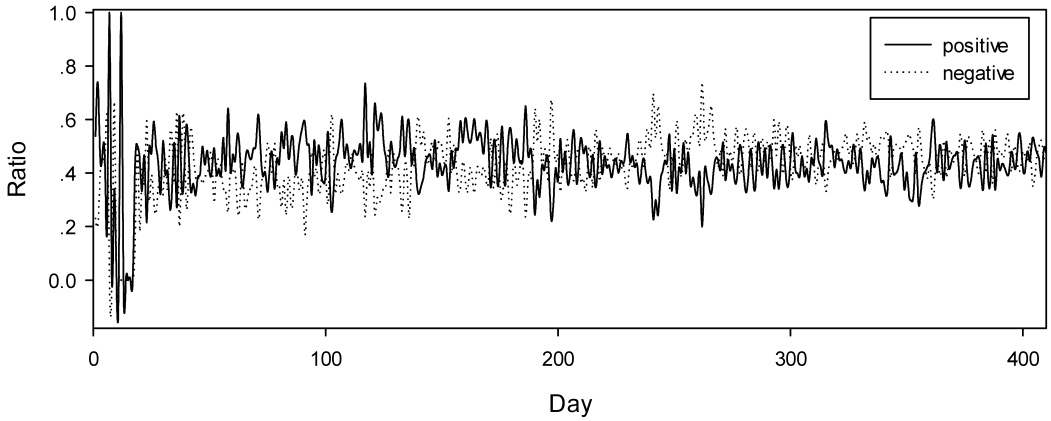

Fig. 7. The ratio of daily positive replies and negative replies in the TCM thread.

3.2. Participants' Attitudes towards TCM

Not only the sentiment of each reply is concerned, the attitude of each reply toward TCM is of more interest, since attitude means one kind of action. After checking a reply is for or against TCM, we find that the attitude toward TCM of each reply is not clear, but the attitude toward TCM of each participant in the discussion is quite clear. We sort the participants by their replies and select 231 participants whose replies are over 10. They have published a total of 81,481 replies, accounting for 91.90% of all the replies. We label their attitudes manually.

Among the 231 participants, 156 support TCM, 63 support abolishing TCM and 12 keep neutral as shown in Figure 8. The number of participants who support TCM is far more than others. Especially, the attitudes of the participants ranked top 10 by replies are listed in Table 2. The number of replies posted by these 10 participants is 57,075, accounting for 64.35% of all the replies. Interestingly, participants listed in Table 2 are half to half by their attitudes toward TCM.

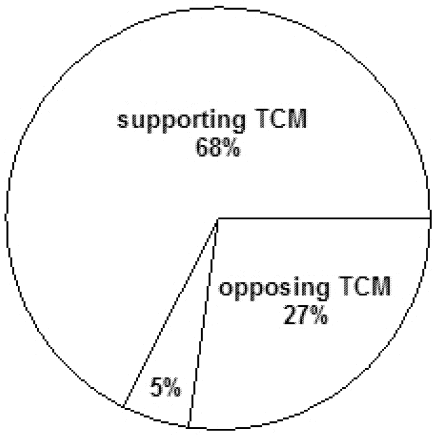

Fig. 8. The attitudes of main participants in the discussion about TCM.

Table 2. The attitudes of participants ranked top 10 by replies

Participants ID	*Replies*	*Attitude toward TCM*
HuoJiGong2012	22,784	Support
EHaiPuXie	9,452	Against
davy1002011	6,907	Support
ShiZhengYi	6,439	Against
sanbenwu	4,029	Support
MuKouLinShiMaBianZi	1,848	Against
djm1000418873	1,638	Against
LaiZheLiKanZheLiYiShi	1,572	Against
JinGu	1,245	Support
LongZhiHunXi	1,161	Support

As each participant's attitude is clear, all the participants are mainly divided into two groups. One group supports abolishing TCM while the other supports TCM. From the point of attitude, the whole discussion also exhibits polarization.

3.3. A Comparison of Sentiment and Attitude Analysis

A cross statistics is made between the sentiment and the attitude based on the 81,481 replies that are posted by the 231 participants with the result listed in Table 3.

Table 3. The relationship between attitude and sentiment

Sentiment *Attitude*	*Positive*	*Neutral*	*Negative*	*Sum*
Support	22,111	4,675	23,475	50,261
Neutral	420	126	275	821
Against	13,310	3,128	13,961	30,399
Sum	35,841	7,929	37,711	81,481

From Table 3, it is found that no matter what kind of the sentiment of a reply, the replies supporting TCM is always far more than the replies supporting abolishing TCM. This is not consistent with the common practice that treats the positive sentiment as the positive attitude and the negative sentiment as the negative attitude.

As a matter of fact, it is necessary to make a distinct difference between sentiment and attitude. From the view of participant, participant's attitude is clear, but his sentiments among his replies are not stable. For example, *HaoJiGong2012* is a firm supporter for TCM, but within all his 22,784 replies, there are 12,116 negative replies and only 9,261 positive replies. From the view of reply, the sentiment is generally definite, but the attitude is not clear. Therefore, it is necessary to distinguish the analytical results of sentiments and attitudes.

4. Conclusion

The recent years have witnessed an unprecedented growth in the volume of text due to the emergence of Web 2.0 platforms and online communities, including blogs, forums, review sites, and social media, which makes online discussions an interesting research field.

This paper focuses on online hot discussions crawled from Tianya Forum and analyzes the discussion about TCM in two aspects: behavior and opinion. It is found that the time interval between replies and the number of participant's replies in the discussion satisfies the power law distribution with the exponent more than 2.0, which is in accord with existing empirical analysis. Online hot discussion has the characteristics of long survival period, many participants and causal language, etc. Moreover, we apply the sentiment analysis method based on sentiment ontology to this thread and manually label the attitudes of the most important 231 participants among 4,890 involved participants. We find that there is a polarization from the view of both sentiment and attitude. Our finding shows that there is a distinct difference between sentiment and attitude, which has not been concerned by most researches.

This paper still has much space for improvement. First, we just analyze one thread of hot discussions about a specific topic for a case study. Since there are so many discussions about the social media, how to combine these separate discussions remains a problem. To tackle this, Floris, *et al* (2013) tried to construct an argument Web and did some experiments [21]. Besides, we recognize that there are some differences between sentiment and attitude, which makes the analysis of online discussions more difficult and complicated. Kim, Tikves *et al*. analyzed the opinion expressed by the Islamic group using SLEP methods and provided a practical way to solve this problem [22].

Acknowledgements. 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&71371107.

References

1. Tang, X. J. 2010. Two Supporting Technologies for Qualitative Meta-synthesis. *Systems Engineering-Theory & Practice*, 30(9): 1593-1606.
2. Tang, X. J. 2009. Qualitative Meta-synthesis Techniques for Analysis of Public Opinions for in-depth Study. In J. Zhou (ed.): *Complex* 2009, Part II, LNICST 5, Springer, pp2338-2353.
3. Luo, B. and Tang, X. J. 2013. Knowledge Vision on Social Network and Guanxi Management Research in Mainland China by the iView Analysis. *Systems Engineering-Theory & Practice*, 33(7): 1661-1671 (In Chinese).
4. Celli, F., Di Lascio, F. M. L., Magnani, M., Pacelli, B., & Rossi, L. 2010. Social Network Data and Practices: the Case of Friendfeed. In *International Conference on Social Computing, Behavioral Modeling and Prediction SBP 2010*. LNCS 6007, Springer, Berlin, pp346-353.
5. Guo, Z., Li, Z., Tu, H., & Li, L. 2012. Characterizing User Behavior in Weibo. In *Proceedings of IEEE Third FTRA International Conference on Mobile, Ubiquitous, and Intelligent Computing (MUSIC2012)*, DOI: 10.1109/MUSIC.2012.18, ISBN: 978-1-4673-1956-0, pp60-65.
6. Guan, W., Gao, H., Yang, M., *et al*. 2013. Hot Social Events on SinaWeibo. *arXiv preprint arXiv:*1304.3898.
7. Cui, L. J., He, H., & Liu, W. 2013. Research on Hot Issues and Evolutionary Trends in Network Forums. *International Journal of u-& e-Service, Science & Technology*, 6(2).
8. Zhao, Y. L. and Tang, X. J. 2013. A Preliminary Research of Pattern of Users' Behavior Based on Tianya Forum. In: Knowledge Creation towards Emergency Management *(proceedings of the 14th International Symposium on Knowledge and Systems Sciences*, Ningbo, October 25-27, 2013 ISBN: 978-4-903092-36-2), Wang S Y, Nakamori Y & Jin W L, eds.), JAIST Press, pp139-145.
9. Pang, B., & Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

10. Xu, L. H., Lin, H. F., Pan, Y., *et al.* 2008. Constructing the Affective Lexicon Ontology. *Journal of the China Society for Scientific and Technical Information*, (2): 180-185 (In Chinese).
11. Hownet. http://www.keenage.com/.
12. Wang, F. Y., Li, X. C., Mao, W. J. and Wang, T. 2013. *Social Computing Methods and Application*. Hangzhou: Zhejiang University Press, 36-39 (In Chinese).
13. Shi, W., Wang, H., and He, S. 2013. Sentiment Analysis of Chinese Microblogging Based on Sentiment Ontology: a Case Study of '7.23 Wenzhou Train Collision'. *Connection Science*, 25(4), 161-178.
14. Introduction of Tianya Forum. http://help.tianya.cn/about/history/2011/06/02/166666.shtml.
15. Zhang, Z. D. and Tang, X. J. 2011. A Preliminary Study of Web Mining for Tianya Forum. In *Proceedings of the 11th Youth Conference on Systems Science and Management Science*, Wuhan: Wuhan University of Science and Engineering Press, 199-204 (In Chinese).
16. Barabasi, A. L. 2005. The Origin of Bursts and Heavy Tails in Human Dynamics. *Nature*, 435(7039): 207-211.
17. Zhu, Y. L., Min, J., *et al*. 2005. Semantic Orientation Computing Based on HowNet. *Journal of Chinese Information Processing*, 20(1): 14-20.
18. Ding, X., & Liu, B. 2007. The Utility of Linguistic Rules in Opinion Mining. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 811-812*.
19. Modak, I. & Mondal, C. 2014. A Study on Sentiment Analysis. *International Journal of Advanced Research in Computer Science & Technology,* 2(2): 284-288.
20. Conover, M.D., *et al.* 2011. Political Polarization on Twitter. In Proceedings of the 5th AAAI International Conference on *Web Blogs and Social Media*. Spain: Barcelona, pp. 237-288. Retrieved from http://truthy.indiana.edu/site_media/pdfs/conover_icwsm2011_polarization.pdf
21. Floris, B., John, L., Mark, S. & Chris, R. 2013. Implementing the Discussion Web. *Communication of the ACM*, 56(10): 66-73.
22. Kim, N., Tikves, S., Wang, Z., Davulcu, H. & Githens-Mazer, J. 2013. Multi-Scale Modeling of Radical and Counter-Radical Islamic Organizations. *HUMAN*, 2(3), pp-182.