# Identifying Suspected Cybermob on Tieba

Shumin Shi[1,2], Xinyu Zhou[1], Meng Zhao[1], Heyan Huang[1,2]

[1.] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[2.] Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications
{bjssm, zxykid, zhaomengBIT, hhy63}@bit.edu.cn

**Abstract** This paper describes an approach to identify suspected cybermob on social media. Many researches involve making predictions of group emotion on Internet (such as quantifying sentiment polarity), but this paper instead focuses on the origin of information diffusion, namely back to its makers and contributors. According our previous findings that have shown, at the level of Tieba's contents, the negative information or emotions spread faster than positive ones, we centre on the maker of negative message in this paper, so-called cybermobs who post aggressive, provocative or insulting remarks on social websites. We explore the different characteristics between suspected cybermobs and general netizens and then extract relative unique features of suspected cybermobs. We construct real system to identify suspected cybermob automatically using machine learning method with above features, including other common features like user/content-based ones. Empirical results show that our approach can detect suspected cybermob correctly and efficiently as we evaluate it with benchmark models, and apply it to actual cases.

**Keywords:** Netizen Identification, Suspected Cybermob, Machine Learning, Support Vector Machine, Social Reviews

## 1 Introduction

Social media on the Internet has become a preponderant channel for the public to express their emotions and share their opinions. By November 2015, the amount of netizens has been up to 688 million in China [1]. Increasingly netizens tend to comment the public affairs on social websites such as Weibo and Tieba with their own language styles, particularly in extremely strong statements.

The public opinion on Internet is the comprehensive expression of individual's beliefs, attitudes, opinions and emotions, which reflect the most netizens' viewpoints facing various emergencies and hot issues. However, cybermobs according to Urban Dictionary[1] and Definithing[2] refer to "*persons acting in cyberspace as to hold someone accountable for a real or imagined misdeed or social faux pas and join together to*

[1] http://zh.urbandictionary.com/define.php?term=cybermob
[2] http://definithing.com/cybermob/

*humiliate or manipulate via the Internet*". And as the main sponsor of the cyber violence or rumors, they always post aggressive, provocative and insulting remarks whenever a public event occurs, which spread negative emotions and falsehoods on the Internet in order to express somewhat extreme private mood or only indulge in flattering by other followers in such an eye-catching way. For example, in the case of the brawl between fans of Sichuan (四川金强) and players of Liaoning (辽宁本溪)(Both are famous basketball teams in CBA) after the 3rd game of CBA-Finals, some cybermobs threw vicious curse-messages on the Weibo of Sichuan Players, which caused the further fierce conflict on social websites among fans from Sichuan Province, Liaoning Province and even the whole country. Besides, instead of paying attention to the truth, cybermobs attack government or people from other region only by virtue of their subjective imagination. Meanwhile, there is a saying "a lie told a thousand times become the truth". The follow-up netizens would rather believe it without verifying the authenticity of the message, which indirectly contribute to the rapid spread of rumors and cause social panic. Those netizens who spread fake information are also suspected cybermobs essentially.

Identifying those rumor-makers or filthy language speakers who could be defined as "suspected cybermobs" in this paper as early as possible is able to make us get ahead of the diffusion of falsity information from the source, help avoid the outbreak of the negative mass mood and cut off the spread of rumors. Aiming at identifying suspected cybermobs, firstly, we automatically collect user information and content information through web-crawler from Baidu Tieba. After collecting user information and content information, and annotating suspected cybermobs, we build a cybermob corpus which is composed by internet users labeled as suspected cybermobs or general netizens with all their postings and their user information. By analyzing the differences of user-based features, content-based features and unique characteristics of suspected cybermobs between suspected group and general ones, we construct a feature set. Finally, we introduce machine-learning methods with features described above to generate several classifiers that are used to identify suspected cybermobs automatically.

The rest of the paper is structured as followed: we will introduce the related work in Section 2. We take cybermobs and general netizens into comparison and collect the features that can distinguish them in Section 3. In Section 4 we train and select a classifier that identify cybermobs correctly with 3 different machine learning methods Then we present our experiment, analyze the results and verify the effectiveness of this approach with actual events. At last we draw a conclusion and point out our future work in Section 5.

## 2 Related Work

### 2.1 Baidu Tieba

Baidu Tieba was established on December 3, 2003 and now is the world's largest Chinese Communicative platform. Here, Tieba is a place on the internet allowing users to do interactive social network site activities. The slogan of Baidu Tieba is "Born for your interest". As of 2014 there have been more than eight million Tieba, mostly created by

fans covering popular stars, films, comics or books. And more than one billion postings have been published in Tieba.

Baidu Tieba ever used to allow anonymous posting which just shows IP address, but now it only allows posting by account, and anonymous posting is not allowed. Users can post with at most 10 pictures and 1 video that can be quoted from certain broadcast websites.

### 2.2 Existing researches

The researches on cybermob are carried out earlier in Sociology Science and generally can be divided into three aspects: the definition, the causes and the guidance of cybermobs. According to the present researches as we know, there isn't an agreement for the definition of cybermob, and here we use the definition of cybermob described in Section 1 in this paper. Studies on the cause of cybermob formation are mainly to measure the negative influence of subjective attitudes and objective factors to the netizens with various methods [2-4]. The guidance of cybermobs is to discuss the possible solution with legislation and education [5-7].

At present, there are not many researches directly on identifying suspected cybermobs. Comparing with the studies on identifying suspected cybermobs, researches on spammers are much deeper. There are mainly 3 kinds of methods for spammer detection [8]. One is based on content feature, For example, Lau RYK et al. [9] utilizes the similarity of contents to find fake reviewers. Liu et al. [10] finds the fake reviews with the help of sentiment classification. Jindal N et al. [11] collects different patterns to automatically find the fake reviews. Another method for spammer detection is based on user feature. Benevento et al. [12] collects 62 features from user's behaviors and detects spammer with them. Zhu [13] proposes a model that automatically selects features based on matrix factorization. User relationship is also used as user feature. Moh et al. [14] gathers user relationship in social media to quantify the user's credibility with which they measure whether a user to be a spammer. Facebook [15] introduces EageRank into this task. The higher the weight, the lower the probability of the candidate user to be a spammer. However, not all cybermobs are born cybermobs, they may have already build their social network before they suddenly become cybermobs. So it may hard to apply these methods into our system. Due to methods based on a specific feature cannot cover all aspects, some methods with integrated multiple specific features are proposed to detect spammers in order to improve the precision [16-18].

Compared with spammers that mechanically publish similar comments, the remarks of suspected cybermobs are full of vulgar speech, region discrimination and criticism of the government. Learning these unique features can improve suspected cybermobs identification system. Suspected cybermobs and general netizens also have differences in user information and content forms. So with user feature and content feature, it will further improves the identification accuracy.

Considering the similarity behaviors between the suspected cybermobs and spammers, especially some suspected cybermobs are directly transformed from spammers, this paper combines methods of traditional methods which are based on user feature and content feature with unique characteristic of suspected cybermobs. After

constructing classifiers with machine learning models, we can automatically identify suspected cybermobs. And with actual events, we verify the method of this paper can provide effective data support for the prediction of group emotion.

## 3 Feature analysis

According to existing researches, general features including user-based features and content-based features are common but effective static features that often used for identification. The number of followers, the number of followees, and whether the avatar is a cartoon picture or a default picture could measure the credibility of general netizens [19]. The ratio that postings with web terminal or mobile terminal is also different in different groups [20]. And even users' location or gender also have an impact on the quality of their postings [21].Besides, suspected cybermobs also have their unique characteristics. Vulgar speech, region discrimination and criticism of the government are nothing new in their postings. In this paper, from labeled corpus, we randomly select 400 labeled as general netizens and 400 labeled as suspected cybermobs each of which is assigned randomly from 1 to 400. We analyze the differences between cybermobs and general netizens from user information, postings' content and unique characteristics of cybermobs with dataset collected.

### 3.1 General feature

**User-based feature (UF).** In the following, cumulative distribution function (CDF) is introduced to analyze the user feature of suspected cybermobs and general netizens, as shown in Figure. 1.

Figure 1(a) and 1(b) analyze the different number of followers and followees between general netizens and suspected cybermobs which illustrates that most the latter ones do not have many followees and they do not care about other users. Figure 1(c) shows the different distributions of average level, which illustrates suspected cybermobs have a higher average level compared to general netizens. As for the number of Tieba postings, suspected cybermobs publish more postings and get more replies than general ones. This maybe because suspected cybermobs always post controversial remarks that lead to the arguments with others. On contrary, a part of netizens like to "div" rather than posting any replies. That's why general netizens post less and why their levels are relatively lower. Analysis on the number of created days from Figure 1(e) indicates that suspected cybermobs tend to create new accounts in a short period. This may be because of their controversial content with which their accounts are easily blocked. While general netizens cherish their accounts and will not change their account frequently.

**Content-based Feature (CF).** Tieba allows users to use functions like posting pictures, emoji, URL links and mention (@) with web terminal or computer terminal. We discuss each of them, and the results are shown in figure 2.

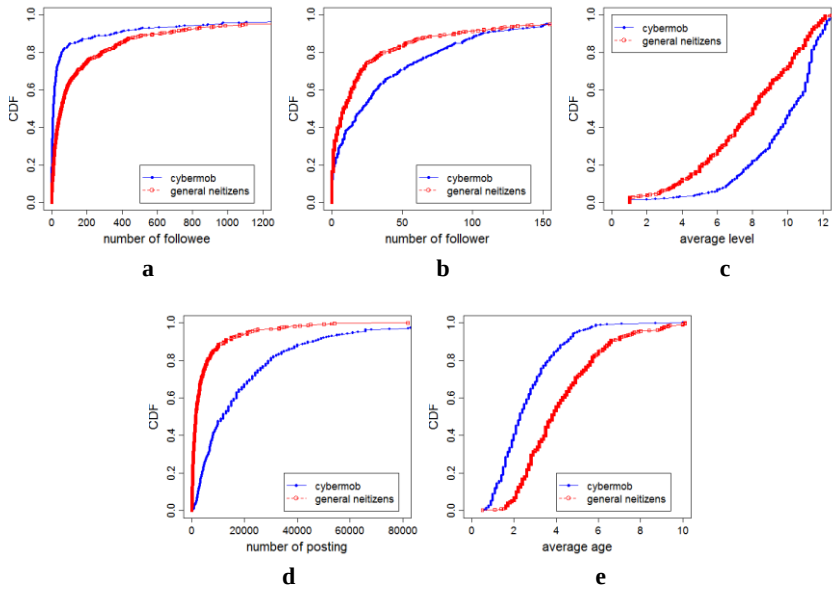


**Fig. 1.** Cumulative distribution function of user-based feature

As we can see in figure 2(a), most users, no matter suspected cybermobs or general netizens do not like to use mention (@).The reason may be that Tieba is not an acquaintance circle, and people do not tend to communicate with each other directly. It is found in figure 2(b) that more than 90% general netizens do not post with emoji. However due to the characteristics of venting emotions, suspected cybermobs post approximately four times on average emoji higher than general ones. As for image shown in figure 2(c), most cybermobs post no more than 0.2 images per posting, while general netizens post at least 0.2 images per posting. Figure 2(d) indicates both suspected cybermobs and general netizens do not like to add URL links into their postings. Maybe it does not work for most time. The ration that people use computer terminal is shown in Figure 2(e). More than half of the general users prefer to post and reply with mobile terminal that is more convenient. While in order to post/reply more and faster, suspected cybermobs tend to post with computer terminal

### 3.2 Unique characteristic of suspected cybermob (UC)

**Vulgar Speech.** In the social media, without supervision, people are free to make remarks. Unscrambling postings in their own views, their comments are full of characteristics of grassroots with current popular internet new vulgar words like “屌丝” (means loser), “叫兽” (means profartssor) which will spread negative emotions and may finally lead to group negative emotions. Therefore, this paper takes the usage of these internet vulgar words into comparison. The results shown in figure 3(a) indicates

that almost every posting of suspected cybermobs contains at least one vulgar word which is many more than that of general ones.

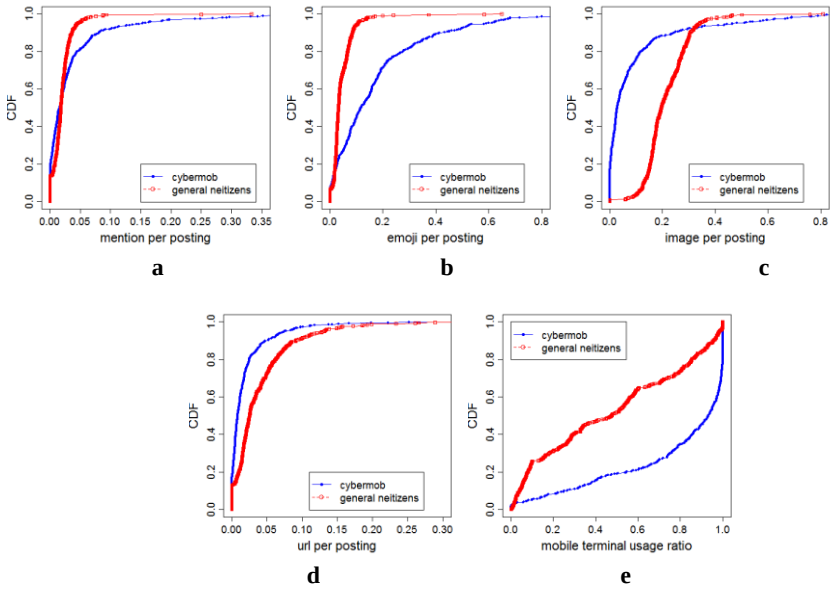


**Fig. 2.** Cumulative distribution function of content-based feature

**Region Discrimination.** There is a prominent region discrimination phenomenon on the internet, namely people due to some native concept of their hometown suffer attacks from netizens who are in other regions, or people take the initiative to attack other areas only because of their "narrow region love". For example, some region discrimination remarks like "XX省人都是些无耻卑贱的狗 (means people in XX province are all shameless dogs)" appear in the comments of the news. Postings from suspected cybermobs often change the topic into abusing each other regardless of the truth, which seriously affect the harmony on the Internet. Therefore, we use these as feature to reveal the cybermobs (as shown in Figure 3(b)). Compared to ordinary users, cybermobs are much more inclined to carry out regional attacks.

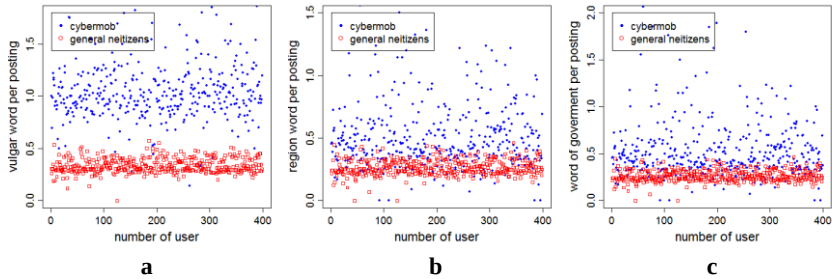


**Fig. 3.** Distribution of characteristic of suspected cybermobs

**Criticism of the Government.** There is another phenomenon that some people prefer to discredit the government on the Internet. No matter what happen, no matter it is right or wrong, they will attract people to the so-called unreasonable policies and blame the government for all mistakes. For example, a medical disputes in Cameroon get to the following comment "这就是中国政府管辖下的医院！（means It is the hospital under the jurisdiction Chinese government)" So aiming at this phenomenon, we first select words associated with government and policy. And compare the frequency between suspected cybermobs and general netizens with negative emotion. The results shown in figure 3(c) illustrate that only cybermobs tend to express negative emotion or directly attack the government and the policies frequently.

# 4 Experiment

## 4.1 Data set

We crawl postings and replies from "NBA" and "dota" in Tieba by the end of April 20, 2015. And we get 3,524,584 postings in total. Then we collect all the users' information and postings with each specific user. We asked two annotators to label the users. The two annotators were requested to judge whether a candidate user to be a suspected cybermob with all his postings and his user information. If there is a disagreement between the two annotators, we ignore this candidate user. We randomly select 400 cybermobs and 400 general netizens as training data. And 150 suspected cybermobs and general users as test data.

## 4.2 Suspected cybermob identification with SVM

Based on the feature above, we introduce SVM (Support Vector Machine) into our suspected cybermob identification system. Figure 4 illustrates the overview of our system: We first collect features as we describe in Section 3 from training data, and construct classifier with machine learning model using these features. After training, the classifier is applied to distinguish whether a candidate user is a general netizen or a suspected cybermob.

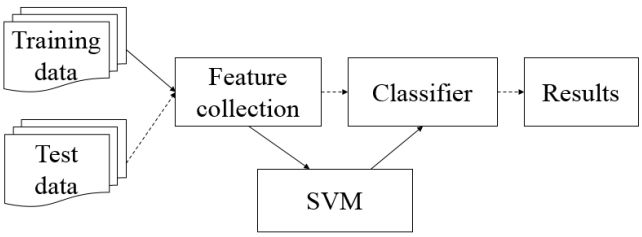


**Fig. 4.** Overview of suspected cybermob identification with SVM

**Support Vector Machine (SVM).** SVM is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis. SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

### 4.3 Benchmark

**Maximum Entropy (ME).** The maximum entropy principle [22] points out that for an unknown situation, we need to predict the probability distribution of a random event, satisfy all known conditions, and uniform distribution of probability for the unknown situation. Because in this case, method of maximum entropy can minimum risks for prediction.

In the process of classification, X is set of feature selection, C is one of the categories, P($c|X$) is a probability with features predicted in category c. Under the restriction of the constraint conditions, the maximum value of the formula (1) is the maximum entropy:

$$H(p) = -\sum_{X,y} P(y|X) \log P(y|X) f(C,t) \tag{1}$$

**Naïve Bayes (NB).** NB is a common technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set using Bayes' theorem:

$$P(c_i|w_1, w_2 \ldots w_n) = \frac{P(w_1|c_i)P(w_2|c_i) \ldots P(w_n|c_i)}{P(w_1)P(w_2) \ldots P(w_n)} \tag{2}$$

All naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature given the class variable. A naive Bayes classifier considers each of these features to contribute independently to the probability.

### 4.4 Results and analysis

We first compare SVM with ME and NB. For each classifier, the same evaluation metrics (precision, recall and F-measure) is calculated for suspected cybermob identification.

As shown in table 1, it is obvious that SVM classifier achieves the best f-measure which indicates SVM could separate training data into two parts with a maximum margin. Besides, NB and ME also achieve a high precision with all features. It shows that

good features including content-based feature, user-based feature and unique characteristics of cybermobs are good contribution to our system.

What's more, we compare the influence of different kinds of features on our suspected cybermob identification system. As shown in Table 1, we achieve the best results with all kinds of features. In addition, compared to the single kind of features, the combination of features can always get better performances, which indicates each kind of feature has a positive effect on the identification of the suspected cybermobs. Compared to user feature and content feature, with unique characteristics, we reach precision at least 88.23%. This result also highlights the unique characteristics that cybermobs tend to vent emotions with vulgar speech, region discrimination and criticisms of the government, with which it is easy to be identified from general netizens.

**Table 1.** Results of different classifiers with different kinds of features

| Feature | SVM | | | ME | | | NB | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F | P | R | F | P | R | F |
| UF | 64.70 | 67.64 | 66.14 | 55.88 | 57.84 | 56.84 | 58.85 | 62.74 | 60.72 |
| CF | 76.47 | 75.49 | 75.98 | 68.62 | 62.57 | 65.55 | 71.56 | 67.76 | 69.61 |
| UC | 94.11 | 93.14 | 93.62 | 88.23 | 83.33 | 85.71 | 90.17 | 87.25 | 88.69 |
| UF+CF | 80.39 | 82.35 | 81.36 | 72.53 | 70.50 | 71.50 | 76.47 | 74.51 | 75.48 |
| CF+UC | 95.10 | 83.14 | 94.11 | 90.17 | 91.18 | 90.67 | 91.18 | 91.18 | 91.18 |
| UF+UC | 97.06 | 96.08 | 96.57 | 94.11 | 93.14 | 93.62 | 96.08 | 92.16 | 94.08 |
| All | 98.04 | 98.04 | 98.04 | 94.11 | 95.10 | 94.60 | 97.06 | 96.06 | 96.57 |

### 4.5 Prototype implementation.

We apply the system into hot spots happened recently in real environment to explore its practical value.

We introduce our system into a hot spot "brawl between fans of Sichuan and players of Liaoning". First, we crawl all the postings from Tieba of "SICHUANJINQIANG (四川金强吧)", "LIAONINGBENXI (辽宁本溪吧)" and "CBA" from March 17th to March 24th 2016. We extract all users who post at least 3 postings in Tieba mentioned above. After collecting user feature, content feature and unique characteristics of suspected cybermobs, we automatically identify whether the users to be suspected cybermobs with our system. 3 annotators are asked to label the users to be suspected cybermobs, and finally we get 103 users (36 cybermobs and 67 general netizens). And the identification results are shown in table 2.

As shown in table 2, we also achieve the best f-measure with all kinds of features. We get higher recall but lower precision with unique characteristics of cybermobs. This maybe because the two teams involved in the brawl are the representatives of Sichuan Province and Liaoning Province, regional vocabularies appear everywhere in posts. And due to the region discrimination, many general netizens are falsely identified as suspected cybermobs. While with user feature and content feature, we get higher precision but lower recall. So in this way, we propose a hypothesis that people play different roles in different events. There may not always be so many suspected cybermobs.

But when an unexpected event related to themselves happens, they may turn into cybermobs to defeat what they support quickly.

**Table 2.** Cybermob identification results of actual events

| Feature | Precision | Recall | F-measure |
|---|---|---|---|
| UF | 86.60 | 36.11 | 50.98 |
| CF | 91.38 | 58.33 | 71.18 |
| UC | 72.00 | 100 | 83.72 |
| UF+CF | 84.00 | 58.33 | 68.86 |
| CF+UC | 72.34 | 94.44 | 81.93 |
| UF+UC | 81.40 | 97.22 | 88.61 |
| All | 86.84 | 91.67 | 89.18 |

### 4.6 Further application

We crawl all the postings by users who are mentioned in Section 4.5 in 2016 except in "SICHUANJINQIANG", "LIAONINGBENXI" and "CBA". In total, we get 3174 postings with these 103 users (36 labeled as general netizens and 67 labeled as suspected cybermobs) and we identify these netizens with their postings in other Tieba using our system again, the results are shown in table 3.

Table 3 illustrate that only 1/3 of those who are automatically recognized as cybermobs with our system before are still identified as cybermobs. And most general netizens identified as general netizens in brawl between Sichuan and Liaoning are still identified as general netizens. Only 2 in 67 who are labeled as general netizens turn into cybermobs because they are involved in other emergencies.

Considering the above analysis, we ensure most cybermobs are not born cybermobs. Nevertheless, when an emergency that may threat to them happens, they will turn into cybermobs in a short period time.

**Table 3.** Results of suspected cybermob identification in different events

| | Suspected Cybermobs | General Netizens |
|---|---|---|
| Suspected Cybermobs in brawl | 12 | 24 |
| General Netizens in brawl | 2 | 65 |

## 5 Conclusion and future work

Aiming at cutting off the diffusion of vulgar-speeches or rumors from source on social websites, we propose an approach for identifying suspected cybermobs who used to be the previous makers or major contributors of the cyber violence or rumors, and always post aggressive, provocative and insulting remarks. We analyze and extract the differences of user-based/content-based feature and unique characteristic of cybermobs firstly. Then we introduce machine-learning method into a practical system based on suspected cybermobs corpus constructed with features above. According to comparison

between different models (SVM, ME and NB, we take ME and NB as benchmark model) and different features restricted, the empirical results show our approach can detect suspected cybermobs correctly and efficiently.

All features play positive roles in our system. However they are all statistic features, we are going to consider applying more cognitive features which are proved to be effective in social psychology into our system to improve the accuracy in next work. Meanwhile, we may miss the golden hours to guide public opinion due to relative hysteresis of social media once events have occurred. So we will take postings-sequence on the Internet into comparison in order that we can identify suspected cybermobs earlier for coming emergency issues.

**Acknowledgments.** We thank reviewers for their constructive comments, and gratefully acknowledge the support of Natural Science Foundation of China (61201352) and the Major State Basic Research Development Program (973 Program) of China (2013CB329606)

## References

1. CNNIC: 36th China Internet Development Statistics Report. China Internet Network Information Center: (2015)
2. Sun W.: Analysis of the causes of "cyber violence". Perspective of Communication Psychology. People Daily (2008).
3. Jun-Xiang L.I.: Reflections on Violence of Public Opinion in Modern Media Environment: Analysis of the Kneeling Incident of Li Yang's Students. Journal of Maoming University. (2008).
4. Tao S.: "Cybermob" with its legal regulation. Theory Research. 109 (2014).
5. Liu X.: Analysis on infringement speech of cybermob. People Daily. China Dominant-journalism Development Center (2007).
6. Hou Z.: Reconstruction of the sense of responsibility to resolve the cyber violence. Democracy and Legal System. 10 (2006).
7. Cao N.. Young children defecation in Hong Kong" the bedizen comments and events "internet mob" phenomenon analysis. Inner Mongolia University. (2015).
8. Qian M., Ke Y.: Overview of Web Spammer Detection. Journal of Software. (2014).
9. Lau R.Y.K., Liao S.Y., Kwok C.W., Xu K., Xia Y., Li Y.: Text mining and probabilistic language modeling for online review spam detection. Acm Transactions on Management Information Systems. 2. 1 (2011).
10. Liu H., Zhao Y., Qin B., Liu T.: Comment Target Extraction and Sentiment Classification. Journal of Chinese Information Processing. (2010).
11. Jindal N., Liu B., Lim E.P.: Finding unusual review patterns using unexpected rules. In: ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October, vol. 1549. (2010)
12. Benevenuto F., Magno G., Rodrigues T., Almeida V.: Detecting spammers on Twitter. In: (2013)
13. Zhu Y., Wang X., Zhong E., Liu N.N., Li H., Yang Q.: Discovering Spammers in Social Networks. In: AAAI Conference on Artificial Intelligence, vol. (2012)

14. Hao S., Syed N.A., Feamster N., Gray A.G., Krasser S.: Detecting Spammers with SNARE: Spatio-temporal Network-level Automatic Reputation Engine. In: Usenix Security Symposium, Montreal, Canada, August 10-14, 2009, Proceedings, vol. 101. (2009)
15. Kincaird J., Edgerank: the secret sauce that makes Facebook's news feed tick. TechCrunch, April (2010)
16. Amleshwaram A.A., Reddy N., Yadav S., Gu G., Yang C.: CATS: Characterizing automation of Twitter spammers. In: Fifth International Conference on Communication Systems and Networks, vol. 1. (2013)
17. Lin C., He J., Zhou Y., Yang X., Chen K., Song L.: Analysis and identification of spamming behaviors in Sina Weibo microblog. In: The Workshop on Social Network Mining & Analysis, vol. 1. (2013)
18. Zheng X., Zeng Z., Chen Z., Yu Y., Rong C.: Detecting spammers on social networks. Neurocomputing. 42. 27 (2015).
19. Morris M.R., Counts S., Roseway A., Hoff A., Schwarz J.: Tweeting is Believing?: Understanding Microblog Credibility Perceptions. Proceedings. 441 (2012).
20. Yang F., Liu Y., Yu X., Yang M.: Automatic detection of rumor on Sina Weibo. In: ACM SIGKDD Workshop on Mining Data Semantics, vol. 1. (2012)
21. Yang J., Counts S., Morris M.R., Hoff A.: Microblog credibility perceptions: comparing the USA and China. In: Conference on Computer Supported Cooperative Work, vol. 575. (2013)
22. Li R., Wang J., Chen X., Tao X., Hu Y.: Using Maximum Entropy Model for Chinese Text Categorization. Journal of Computer Research & Development. 42. 578 (2005).