Detecting Spammers on Social Networks Based on a Hybrid Model

Guangxia Xu*, Jin Qi†, Deling Huang‡ and Mahmoud Daneshmand§
*School of Software Engineering, CQUPT, Chongqing 400065, China
E-mail: xugx@cqupt.edu.cn
*Information and Communication Engineering postdoctoral research station
Chongqing University, Chongqing 400044, China
†School of Software Engineering, CQUPT, Chongqing 400065, China
E-mail: 1359918905@qq.com
‡School of Software Engineering, CQUPT, Chongqing 400065, China
E-mail: huangdl@cqupt.edu.cn
§School of Business, Stevens Institute of Technology, Hoboken NJ 07030-5991, USA
E-mail: mdaneshm@stevens.edu

***Abstract*—The prosperity of social networks provides users with convenient communication but also attracts a large number of spammers. To solve this problem, this paper combines supervised learning and unsupervised learning algorithms, and proposes a novel hybrid model based on OPTICS and SVM. First, we collected a dataset from Sina Weibo including 10,000 users and 134,188 messages; then extracted the content based features and user behavior based features from the dataset; afterwards, we applied the features into the hybrid model to establish the classification model. The experiment shows that the proposed approach is capable of detecting spammers effectively with 87.6% spammers and 94.7% legitimate users correctly classified.**

***Keywords*-social network;spammer;hybrid model;**

I. Introduction

Social network, also known as social network service, social media networks or social network sites, is the network platform on which people who have common interests, behaviors and backgrounds establish social relationships [1]. As the Internet industry structure and user behavior is changing along with the rapid development of the Internet, social networks flourish around the world and become a new growth momentum of the Internet industry. The popular social networks such as Twitter, FaceBook and Sina Weibo have a large number of users. The monthly active users of FaceBook have exceeded 1.6 billion and that of Sina Weibo in China have reached 216 million as of March 2016 [2-3].According to the statistics issued by the authoritative agency Alexa, which provides leading web site ranking around the world, more than half of the top 20 sites are available for social network service [4]. Social networks have become important communication platforms in people's life.

Social networks have the characteristics of diversification, popularization, high real-time and interactivity and provides people with convenient communication. However, it also provides a large quantity of opportunities for spammers, which spread malicious messages such as advertising, pornography, phishing sites, fake news and so on. These messages affect the normal access to information of the user and threaten the legitimate users privacy information and account security, which adversely affects the user experience. At the same time, these malicious messages lead to serious costs of network resources, interfere with the normal data mining and analysis, and increase the operational burden of social network. Besides, there are also a number of spammers making profits through marketing promotion behaviors including malicious likes, comments, votes and reposts, which severely harm the credibility evaluation system of the social network and user's trust relationship.

The traditional techniques of spammer detection are mainly divided into the method of feature analysis and machine learning which can be divided into supervised learning and unsupervised learning, and the method of community network or graph. Using supervised learning algorithm relies on costly human inspection of training dataset in advance for building classifiers, but the spammers usually adapt their strategies to bypass the classifiers. Although using the unsupervised learning algorithm or community network does not need labeled dataset to train classifiers, the accuracy of them is low. As a result, this paper proposes a novel detection method based on a hybrid model combining ordering points to identify the clustering structure (OPTICS) and support vector machines (SVM). The method contains the following three steps: firstly, crawl a dataset from Sina Weibo containing 10,000 users and 134,188 messages posted by them; secondly, select content based features and behavior based features of users which have discriminative power to distinguish spammers and legitimate user; lastly, establish a hybrid classification model combining OPTICS and SVM to detect spammers.

The rest of the paper is organized as follows. Section 2 summarizes the related works about spammer detection; Section 3 introduces how we collect data and extract features, and describes the hybrid classification model based on OP-

978-1-4673-9005-7/16/$31.00 ©2016 IEEE

TICS and SVM; in Section 4, the experiment is conducted to verify the effectiveness of hybrid model; finally, Section 5 concludes the paper.

II. Related Work

According to the harm caused by spammers on social network, the existing researches of spammer detection are mainly divided into the method of feature analysis and machine learning which can be divided into supervised learning and unsupervised learning, and the method of community network or graph. The method of feature analysis and supervised learning is most widely used, of which the research is relatively mature. In the first study applying machine learning to spammer detection [5], the features of tweets and user behaviors were used to train SVM classifier to distinguish spammer and legitimate user. Similarly, Zheng et al. [6] crawled a big dataset from Sina Weibo including 30,116 user's personal information and more than 1600 million messages to train the SVM classifier. Besides, different feature sets have different contributions to spammer detection. Ahmed et al. [7] adopted 14 kinds of general statistical features and three classification algorithms including simple Bias, Jrip, and J48 to detect spammers respectively. At the same time, they carried out a series of experiments to calculate the contributions of features to spammer detection by deleting one feature every time. However, the spammers usually adapt their strategies to bypass the methods above. To solve this problem, Lee et al. [8] deployed social honeypots on social network to capture spammers and adopted the best classification method of decorate algorithm to detect unknown spammers through comparing the performance of the 10 classifiers.

For the study of feature analysis and unsupervised learning, Miller et al. [9] extracted 12 user information features and 95 content features from Twitter and used a data stream clustering algorithm combining DenStream and StreamKM++ to cluster. The method generates legitimate user clusters and the rest of the users outside the clusters are spammers. Cao et al. [10] found that spammers on social networks often exhibit a loose synchronous behavior. Therefore, they clustered users based on the similarity of user behavior to discover spammers which expose continuous similar behaviors over a period of time.

In addition to the methods based on machine learning algorithm, there is another kind of researches that establish spammer detection model based on user's relationship network. Yang et al. [11] found that spammers usually form small networks and three types of users have close connections with spammers. Bhat et al. [12] found that legitimate user groups on social network have their own personality characteristics, and thus proposed a classification model based on the characteristics of community framework to distinguish spammers and legitimate users.

Although the methods of supervised learning above are mature and have high detection accuracy, they need to construct a labeled dataset relying on costly manual inspection to train the classifiers. However, since the spammers usually adapt their strategies to bypass the classifiers which lead to the labeled dataset ineffective, these methods of supervised learning need to continuously construct the labeled training data and classifiers, consuming quantities of labor costs and training costs. Similarly, even though using unsupervised learning algorithms or community network does not need labeled dataset, the accuracy of them are relatively low. As a consequence, this paper proposes a novel detection method based on OPTICS and SVM hybrid model which combines supervised learning and unsupervised learning algorithms to detect spammers from different levels.

III. Spammer Detection in Sina Weibo

A. Overview of Spammer Detection Framework

In this part, the framework of spammer detection method based on OPTICS and SVM hybrid model proposed in this paper is briefly introduced. Firstly, collect data from Sina Weibo including user information and messages posted by them through web crawler; secondly, extract content based features and behavior based features of users which have discriminative power to distinguish spammers and legitimate user by analyzing the cumulative distribution function (CDF) curves of the features; finally, input the unlabeled feature vectors to the hybrid classification model combining OPTICS and SVM to detect spammers. The overall framework of the method is shown in Fig. 1.

B. Crawling Sina Weibo

Since the existing API of Sina Weibo does not open completely, we wrote a web crawler program to crawl user data. At first we crawled 10,000 user's personal information and captured the latest 50 messages posted by them and the number of likes, comments, reposts and post time of the messages in May 2016. Thus, we collected 10,000 users and 134,188 messages in total. Afterwords, in order to evaluate the performance of our spammer detection model accurately, we manually labeled the crawled users of which classes is spammer or legitimate user using the method that is similar to the literature [6]. The method identified three volunteers. They analyzed each collected user manually based on the recent messages and gave a vote independently. Then the class of majority votes is the label of the user. However, because the user labeling method depends on artificial judgment, it might result in inevitable human error. As a result, we ignore those users of which the class is fuzzy. Lastly, 1,311 spammers and 4,189 legitimate users are labeled.

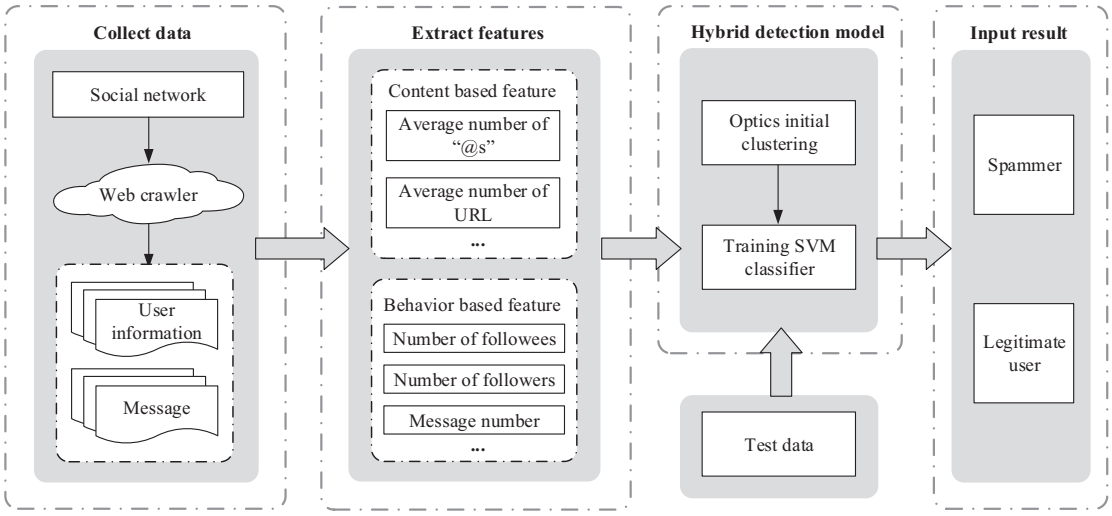


Figure 1. Overall Framework of Spammer Detection.

C. Content Based Features Analysis

In this paper, we divide the user features into content based features and user behavior based features. In this part, we analyze the content based features through CDF curves. There is a great difference between the messages posted by spammers and legitimate users. Unlike legitimate users, spammers usually post advertising, pornography, phishing sites, fake news and other malicious messages. As a consequence, extracting features from the content of messages can effectively detect spammers. In this paper, we define the features that calculated through content of messages as the content based features, such as the characteristics of "@", topic, URL, message similarity and so on.

Usually, in Sina Weibo users can use "@username" when posting messages to refer to others. Spammers may exploit this to mention other users who are not their followers to spread malicious messages. And users in Sina Weibo utilize "#topic#" to participate in the discussion of a certain topic. The "#topic#" is similar to a tag assigned to a message and the most popular "#topic#" that appears in messages becomes trending topic. The message of a certain user containing "#topic#" can not only be seen by the followers, but also be posted in the bulletin board to be seen by more users. Therefore, spammers would be more likely to post messages containing "#topic#" that are unrelated with the topic to increase the possibility to be searched and displayed. Besides, due to the limit of 140 characters per message, Sina Weibo provides a link-shortening service for users to reduce the length of their URLs, thus it is difficult to decide whether the URLs are malicious or not. Spammers abuse this service and post more messages containing malicious URLs to deceive users, of which the destinations are shopping, phishing and pornography sites. At the same time, spammers may post messages generated by certain content of templates, resulting in the high message similarity. The average message similarity of user is calculated as follows:

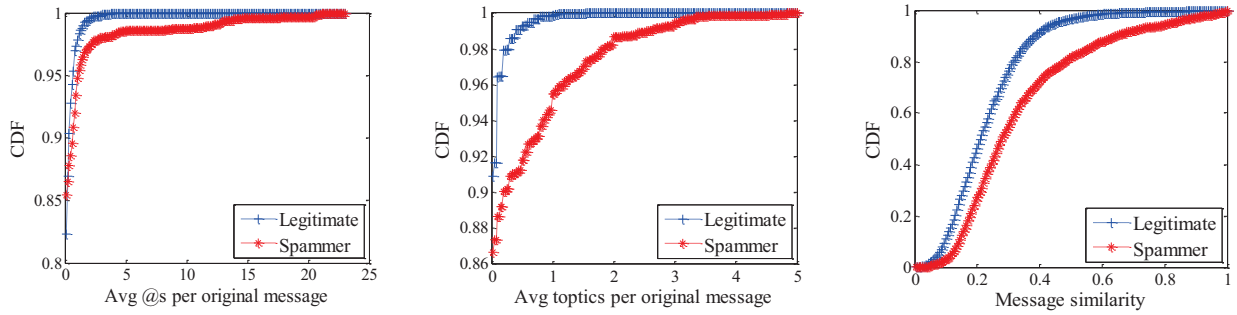
$$\sum_{a,b \in T} \frac{S(a,b)}{N} \tag{1}$$

T is the set of message pairs posted by a user,N is the number of message pairs, and $S(a,b)$ is the similarity of message a and b. In order to compute $S(a,b)$, firstly we delete the "@s", "#s", URLs, emoticons and stop words in the messages, and then transform the plain text into vector space model composed of feature words through tokenizer. Then we use the theory of term frequency-inverse document frequency (TF-IDF) to calculate the weights for each feature word and utilize the standard cosine similarity over the vector space model to get the message similarity:

$$S(a,b) = \frac{\overrightarrow{V}(a) \cdot \overrightarrow{V}(b)}{\left|\overrightarrow{V}(a)\right| \left|\overrightarrow{V}(b)\right|} \tag{2}$$

$\overrightarrow{V}(a)$ and $\overrightarrow{V}(b)$ represent the vector space model of message a and message b respectively.

Based on the previous collected data, we compute the content based statistical features, including the average number of "@s", topics, URLs and pictures, the average number of "@s" and topics per original message, the average message length and similarity. Then we draw the CDF curves of those features to study their discriminative power for spammers and legitimate user. As shown in Fig. 2, we analyze three content based features in detail. It can be clearly noted from Fig. 2(a) that the average number of "@s" in the original messages posted by spammers is



(a) Average Number of "@s" per Original Message

(b) Average Number of Topics per Original Message

(c) Message Similarity

Figure 2. Cumulative Distribution Functions of Three Content Based Features.

higher in comparison with legitimate users. Similarly in Fig. 2(b), the average number of topics in the original messages posted by spammers is higher as well. While the majority of legitimate users normally have not more than one topic in their original message, some spammers use many topics in their original message to increase the possibility to be searched and displayed. Fig. 2(c) shows that the message similarity of spammer is higher than legitimate user.

D. User Behavior Based Features Analysis

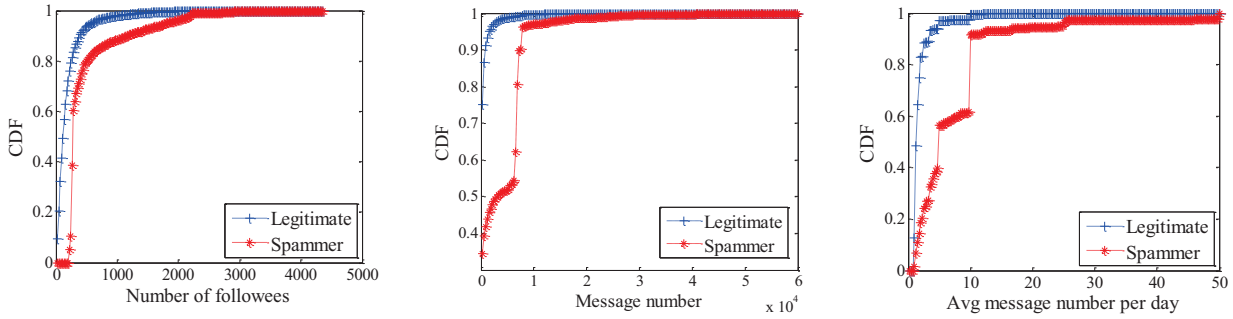
The behaviors of users on social networks are different due to their different purposes. Because of the simple purpose of entertainment, browsing news and other normal requirements, the behaviors of legitimate users are relatively unitary. However, due to the purpose of making profits, spammers would show some abnormal performance, such as following a lot of users in a short period of time, improving the frequency of message posting and so on. In this paper, we define the user information, the average number of likes, reposts, comments of messages and the frequency of message posting as the user behavior based features.

Through the previous collected data, we calculate the user behavior based statistical features, including sex, the number of followers, the number of followees, the fraction of followers per followees, the number of messages, user level, user authentication, the average number of messages per day and the average number of likes, reposts, comments of messages. Then we analyze these features through CDF curves. Fig. 3 displays three user behavior based features. It can be clearly noted from Fig. 3(a) that spammers follow more users in comparison with legitimate user so as to be followed back. Fig. 3(b) indicates that the number of messages of most legitimate users is below 10,000 and that of some spammers is far beyond 10,000. They are probably marketing users which repost a multitude of messages to improve the attention of an activity or a certain user. Similarly, in Fig. 3(c), it is can be found out that spammers post more messages than legitimate users every day.

E. A Hybrid Classification Model Based on OPTICS and SVM

OPTICS is a density based clustering algorithm, which improves the algorithm of density-based spatial clustering of applications with noise (DBSCAN) and is not sensitive to the input parameters of neighborhood radius ε and neighborhood density threshold $MinPts$. The algorithm outputs a cluster ordering of the dataset and stores the core-distance $cdist$ and reachability-distance $rdist$ of each object instead of generating clusters directly [13]. The reachability-distance is related to the space density of the object. If the space density is large, the reachability-distance of the object from its adjacent object is small. In order to make the cluster expand as far as possible to the dense space of the data, the OPTICS algorithm selects the object of which reachability-distance is minimum to extend. Therefore, OPTICS outputs the cluster ordering by maintaining a table named Order-Seeds of which the objects are sorted in ascending order by reachability-distances. The cluster ordering represents the density based cluster structure of all the objects and is equals to a DBSCAN cluster result of a wide range of parameter settings. Given a parameter of neighborhood radius ε' that is less than or equal to ε , a cluster result that is similar to DBSCAN can be extracted from the cluster ordering.

The basic idea of SVM is that in the case of linear separable and linear non-separable, it is able to construct a hyper plane that divides the training dataset correctly and make the geometric interval between the two classes largest [14]. In the case of linear non-separable, it uses kernel functions to divide data. The kernel function is a nonlinear mapping method, which maps the training data in original space to a high dimensional feature space to make the data linearly separable in the high dimension space. In this paper, we adopt polynomial kernel function.



(a) Number of Followees (b) Number of Messages (c) Average Number of Messages per Day

Figure 3. Cumulative Distribution Functions of Three User Behavior Based Features.

The classification accuracy of SVM is high, but its result highly depends on the selected training data and requires a large number of high quality labeled data. Although using OPTICS to detect spammers is fast and does not need labeled dataset, the accuracy is relatively low. The hybrid classification model proposed in this paper combines the two algorithms. In the first, it uses OPTICS to generate cluster ordering and extracts initial clusters through the cluster ordering. Then, it acquires some reliable learning samples from the initial clusters to train SVM classifier. The classification process of the hybrid model is divided into three parts:

1) Initial cluster by OPTICS. Input the neighborhood radius ε , the neighborhood density threshold $MinPts$ and the dataset $X = \{x_1, x_2, \cdots, x_n\}$ into OPTICS to get a ordered list of X , the core-distance $cdist$ and reachability-distance $rdist$ of each object. Given a parameter of neighborhood radius ε' $(\varepsilon' \leq \varepsilon)$, a cluster result is generated by the Cluster Extracting algorithm of OPTICS.
2) Selecting training samples. The initial class labels have been generated by OPTCIS in the previous step. In order to select the samples of dense part from spammer cluster and legitimate user cluster respectively, a threshold of reachability-distance $rdist'$ is given to select samples of which $rdist$ is less than $rdist'$.
3) Re-classification by SVM. The initial labeled samples from the previous step are used to train SVM classifier. Then use the trained SVM classifier to re-classify the original dataset and get new classification result.

IV. Experiments

A. Evaluation Metrics

To evaluate the effectiveness of our classification model, we use the standard information retrieval metrics of precision (P), recall (R) and F-Measure (F) [15] and illustrate them by the confusion matrix of Table I. TP refers to the number of spammers correctly classified; FN is the number of spammers mistakenly predicted as legitimate users; FP refers to the number of legitimate users misclassified as spammers; TN represents the number of legitimate users correctly classified.

Table I. Example of Confusion Matrix

		Predicted	
		Spammer	Legitimate user
True	Spammer	TP	FN
	Legitimate user	FP	TN

The precision of spammer is the ratio of the number of correctly classified spammers to the total number of predicted spammers and is defined as $P_{spammer} = TP/(TP + FP)$. The recall of spammer is the ratio of the number of correctly classified spammers to the total number of true spammers and is expressed as $R_{spammer} = TP/(TP + FN)$. F-Measure refers to the harmonic mean of precision and recall and is defined as $F = 2PR/(P + R)$. It can evaluate the performance of classifiers more comprehensive and precisely.

Meanwhile, the evaluation of clustering algorithms is usually realized by calculating the corresponding degree of cluster labels and the actual class labels. In this paper, we adopt the classification-oriented evaluation [16]. By letting the cluster labels equal to the predicted class label, the precision, recall and F-Measure that are similar to the classification metrics are used to assess clustering algorithms.

B. Experiment Result and Comparison

Through feature analysis in this paper, we use 18 features listed in the following: the average number of "@s", topics, URLs and pictures, the average number of "@s" and topics per original message, the average message length, message

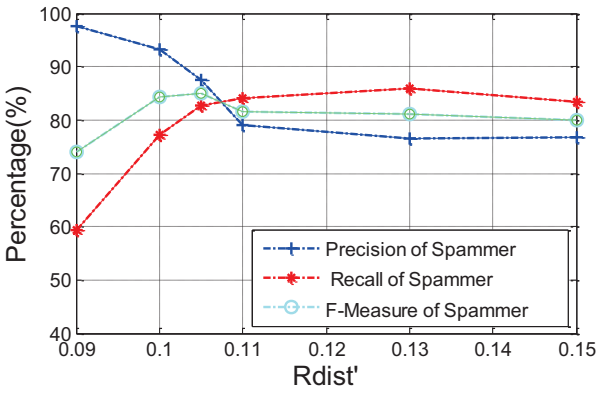


Figure 4. Influence of the Variation of Rdist' on the Classification Results.

Table II. Confusion Matrix of Hybrid Model

		Predicted	
		Spammer	Legitimate user
True	Spammer	87.6%	5.3%
	Legitimate user	12.4%	94.7%

Table III. Classification Evaluation of Hybrid Model

	P	R	F
Spammer	0.876	0.828	0.851
Legitimate user	0.947	0.963	0.955

similarity, the number of followers, the number of followees, the fraction of followers per followees, the number of messages, user level, user authentication, the average number of messages per day and the average number of likes, reposts, comments.

First of all, the OPTICS algorithm is used on the original dataset to generate initial clusters of which the parameter ε is set 0.18 and $MinPts$ is set 19. Then based on the Cluster Extracting algorithm of OPTICS, three clusters are generated: legitimate user cluster of number 0, legitimate user cluster of number 1 and spammer cluster of number 2. Afterwards, the threshold of reachability-distance $rdist'$ is given to select reliable samples from the initial clusters. The value of $rdist'$ has a great influence on the classification results. It is concluded from experiment that when $rdist'$ varies between 0.09 and 0.15, the recall and F-Measure of legitimate user are both above 92%, the precision of legitimate user is above 89%, but these metrics of spammers have a big range of variation. Therefore, we only analyze the influence of the variation of $rdist'$ on the classification results of spammers. As shown in Fig. 4, when $rdist'$ varies between 0.09 and 0.15, the precision of spammer exhibits a gradual upward trend, the recall exhibits a gradual downward trend and the F-Measure increases first, reaches the maximum at 0.105, and then gradually decreases.

That is because when $rdist'$ is low, the space density of selected learning samples is high and the number of samples having false initial labels in the training samples is small, thus the precision is high; at the same time, the number of learning samples selected from initial clusters is small which result in that the SVM classifier learns less knowledge from the training samples, thus the recall is low. In the same way, when $rdist'$ is high, the precision is low and the recall is high. At last, the $rdist'$ is set 0.105.

After training samples selection and SVM classifier training, the trained SVM classifier is used to update the initial class labels. Table II describes the confusion matrix of our hybrid model. It shows that our proposed model is efficient, with 87.6% spammers and 94.7% legitimate user correctly classified. Table III illustrates the values of classification evaluation including precision, recall, and F-Measure of spammer and legitimate user.

Besides, we also compare the proposed approach with OPTICS and SVM. As is shown in Table IV, it is obvious that the OPTICS algorithm exhibits worst result and the SVM classifier achieves best accuracy. The detection result of our proposed hybrid model is between them and achieves great improvement than the former without labeled dataset.

V. Conclusion

In order to solve the problems that the detection method of unsupervised learning has low accuracy and supervised learning needs a large number of manually labeled dataset, a spammer detection method based on the hybrid classification model combining OPTICS and SVM is proposed in this paper. The experiment on the dataset crawled from Sina Weibo shows that the method is efficient and does not need to label the dataset manually.

However, there are still some inadequacies in this paper. In order to evaluate the proposed approach, we still label the collected dataset by human inspection, which costs a lot

Table IV. Comparison between Three Detection Methods

	Spammer			Legitimate user		
	P	R	F	P	R	F
OPTICS	0.609	0.510	0.555	0.946	0.864	0.894
Hybrid model	0.876	0.828	0.851	0.947	0.963	0.955
SVM	0.939	0.872	0.904	0.961	0.982	0.971

of labor and time. Thus we elected to use a smaller dataset which could be manually labeled. However, in the practical application of large scale datasets the proposed method takes a long time to extract features and train model, which cannot meet the needs of real-time detection. How to shorten the training time and ensure the accuracy is the problem to be solved in the future.

ACKNOWLEDGMENT

The authors acknowledge supports from the National Natural Science Foundation (Grant No. 61309032, Grant No. 61272400), China Postdoctoral Science Foundation (Grant No. 2014M562282), the Project Postdoctoral Supported in Chongqing (Grant No. Xm2014039), the Wenfeng Leading Top Talent Project, the New Research Area Development Programme (Grant No. A2015-44), the Science and Technology Research Project of Chongqing Municipal Education Committee (Grant No. KJ1400422), Common Key Technology Innovation of Important Industry by Chongqing Science and Technology Commission (Grant No. CSTC2015ZDCY-ZTZX40001), Collaborative Innovation Center for Information Communication Technology (Grant No. 002), the Science and Technology Research Project of Chongqing Municipal Education Committee (Grant No. KJ1500441), the Science and Technology Research Project of Chongqing Municipal Education Committee (Grant No. KJ1400431).

REFERENCES

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.

[2] Stats of facebook. [Online]. Available: http://newsroom.fb.com/company-info/

[3] Earning of sina weibo. [Online]. Available: http://earnings.card.weibo.com/2016Q1/

[4] Alex top 500 global sites. [Online]. Available: http://www.alexa.com/topsites/global

[5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, 2010, p. 12.

[6] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 42, no. 1, pp. 27–34, 2015.

[7] F. Ahmed and M. Abulaish, "A generic statistical approach for spam detection in online social networks," *Computer Communications*, vol. 36, no. 1011, pp. 1120–1129, 2013.

[8] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July*, 2010, pp. 435–442.

[9] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Information Sciences*, vol. 260, no. 1, p. 6473, 2014.

[10] Q. Cao, X. Yang, J. Yu, and C. Palow, "Uncovering large groups of active malicious accounts in online social networks," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 477–488.

[11] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 71–80.

[12] S. Y. Bhat and M. Abulaish, "Community-based features for identifying spammers in online social networks," in *Ieee/acm International Conference on Advances in Social Networks Analysis and Mining Asonam*, 2013, pp. 100–107.

[13] M. Ankerst, "Optics: ordering points to identify the clustering structure," in *SIGMOD 1999, Proceedings ACM SIGMOD International Conference on Management of Data, June 1-3, 1999, Philadelphia, Pennsylvania, Usa*, 1999, pp. 49–60.

[14] C. Cortes and V. Vapnik, "Support vector network," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[15] Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, vol. 1, no. 1, pp. 69–90, 1999.

[16] T. Pangning, S. M, and K. V, *Introduction to Data Mining*. Posts and Telecommunications Press, 2011.