Detecting Internet Hidden Paid Posters Based on Group and Individual Characteristics

Xiang Wang[1(✉)], Bin Zhou[1,2], Yan Jia[1,2], and Shasha Li[1]

[1] School of Computer, National University of Defense Technology, Changsha, China
{xiangwangcn,binzhou,yanjia,shashali}@nudt.edu.cn
[2] State Key Laboratory of High Performance Computing,
National University of Defense Technology, Changsha 410073, China

Abstract. Online social networks are popular communication tools for billions of users. Unfortunately, they are also effective tools for hidden paid posters (or Internet water army in some literatures) to propagate spam or mendacious messages. Paid posters are typically organized in groups to post with specific purposes and have flooded the communities of microblogging websites. Typical traditional methods only utilize individual characteristics in detecting them. In this paper, we study the group characteristics of paid posters and find that group characteristics are also very important in detecting them comparing to individual characteristics. We construct a classifier based on both the individual and group characteristics to detect paid posters. Extensive experiments show that our method is better than existing methods.

Keywords: Paid posters · Internet water army · Microblogging · Social network

1 Introduction

Nowadays, social networks like Twitter, SINA Weibo and Facebook are becoming popular information sources for billions of people. Due to the ease of forwarding messages, information can disseminate to a large number of interested people via their social network. For example, if a user posts a tweet in Twitter, all its followers can read the tweet immediately. Some users like famous people even have millions of followers. As one of the major social networks, microblogging differs from a traditional blog and allows users to exchange small elements of content such as short sentences, individual images, or video links. There are several famous microblogging platforms like Twitter, SINA Weibo and Yammer. Users can post about topics ranging from the daily chats to the thematic like national policies. The microblogging platforms have significantly influenced people's daily life and brought considerable opportunities to business.

Paid posters [8] are typical employed to promulgate spam or mendacious information to increase normal users' awareness of their targets in a campaign. If there are large number of mendacious messages, normal users can not know the

J. Wang et al. (Eds.): WISE 2015, Part II, LNCS 9419, pp. 109–123, 2015.
DOI: 10.1007/978-3-319-26187-4_8

actual state of affairs. This maybe lead some bad consequences if large number of normal users are misled. For example, the CEO of Smartisan Technology corporation named Luo Yong-hao announced that their new product "Smartisan T1" was attacked by paid posters and publicly offered a reword of CNY 200,000 to find the paid posters[1]. There are many slanderous comments for their new product and the sales of the product are decreased. To reduce the negative effect, it's crucial for us to detect paid posters.

Paid posters are different from traditional spammers. First, paid posters are typical a group of users with group characteristics while spammers (except opinion spam) are usually considered to be individuals. Second, some paid poster groups go far away than posting spam message, the behaviors of them sometimes can hurt others. Third, paid posters are either controlled by a program through platform API or human beings. They are different from Twitter bot [9] which is a program used to produce automated posts or to automatically follow Twitter users. As they can also be human beings which are more covert and complex than Twitter bot. Fourth, paid posters are more covert than spammers. They are normal users at ordinary times, but they become paid posters when they try to promote a campaign. Even Some famous users with high influence can be paid to be paid posters temporarily when they are needed in a promoting campaign. Opinion spam is a kind of paid posters [16,20], but existing researches focused on detect them in electronic-commerce websites like Amason.com and hotel booking website TripAdvisor, rather than social network platforms like microblogging websites.

Spammers have been appearing in many applications like blogs [17,24], email [2,6], Web search engine [12] and videos websites [3,5]. There are a large amount of methods which have been proposed to detect them [11,15] in these platforms. They mainly employ individual statistical characteristics for detecting spam. Our study finds that group characteristics are also important for detecting paid posters. Traditional methods which only utilize individual statistical characteristics are not good enough for detecting paid posters. For example, in a promotional campaign to promote an URL, many paid posters retweet the advertising tweet to their communities to make large number of users see it. Typically most of them do not follow the author of the advertising tweet, so it is important to use group characteristic "retweeting without following" to detect paid posters.

In this paper, we study six group characteristics for detecting paid posters. The group characteristics are discussed in Sect. 3. Some individual characteristics used in traditional spam detecting methods are also utilized in our method. User influence which is calculated by users' multi-relational networks [10] is employed to detect paid posters. We employ the SVM model to combine the individual characteristics and group characteristics to detect paid posters. Experimental results on three real datasets show that our method is better than existing methods.

The main contributions of this paper can be summarized as follows:

[1] http://digi.163.com/14/0919/15/A6H2KS8H00162OUT.html.

- We study several useful group characteristics for detecting paid posters and find that group characteristics are also very important in detecting paid posters comparing to traditional individual features.
- We propose a method named "IGCSVM" combining both user's individual and group characteristics for detecting paid posters.
- Extensive experiments have been done on three real datasets of SINA Weibo. Experimental results show that our IGCSVM method is more effective than existing methods in detecting paid posters.

The rest of this paper is organized as follows: Sect. 2 discusses some important related works. Section 3 introduces our method for detecting paid posters. Experimental results are shown in Sect. 4. Finally, conclusion and future work are provided in Sect. 5.

2 Related Works

Spammers have been appearing in a lot of applications, such as blogs [17,24], email [2,6], Web search engine [12] and videos [3,5]. And there are a large amount of methods which have been proposed to detect them [11,15]. Zhang et al. [26] analyze the characteristics of the spam users in two campaigns in Twitter. They explored the mention network to find the characteristics of outdegree and indegree, neighborhood connectivity and burstiness in order to find their relationships with spam users. They also analysis the online social network to get the features of followers/friends and response time. They try to find useful features for spam detection. They also investigate the benefit-cost analysis of spammers based on epidemic model. Yang et al. [27] presented a case study of analyzing inner social relationships of criminal users and proposed a new algorithm named Mr.SPA to detect users that have close relationship with criminal users. They also designed an algorithm named CIA to detect more criminal users based on a seed set by analyzing the social and semantic relationships among users. Gao et al. [13] proposed a method to detect malicious users and posts based on URL and text clustering. They also analysis the characteristics of the malicious users and posts. Thomas et al. [23] characterized the behaviors of 1.1 million spammers on Twitter by analyzing the text of the tweets sent by the suspended users. They also found there was a market providing spam users services. They also explored five spam campaigns and find the tools employed by spammers and the approaches they used in spam activities. Lee et al. [18] analyzed the profile features of spammers and developed a classifier to classify spam users to different categories: promoters, legitimate users and so on. Grier et al. [14] studied spam on Twitter and found that clickthrough rate of spam URLs was much lower than email. The analyze also showed that 84 % spam users are organized by few number of controllers. M. McCord and M. Chuah [19] studied user based and content based features and find that they are different between spammers and legitimate users. They also utilize the features for detecting spammers. Chu et al. [9] build a classifier to determine an account to be a human, bot or cyborg.

There are also some researches about paid posters. Opinion spam is a kind of paid poster. Jindal and Liu [16] found that opinion spam is widespread and in electronic-commerce websites. They trained their models using features like review text, reviewer and product to detect duplicate opinions in Amazon.com. Ott et al. [20] proposed a n-gram based text categorization to detect deceptive opinion spam in hotel booking website TripAdvisor. Chen et al. [8] investigated the behavioral pattern of paid posters and designed a detection mechanism to identify potential paid posters based on user comments in social network. We utilize not only user comments but also user posts, user social friendships and group characteristics for detecting paid posters in this paper. Wang et al. [25] studied five features for detecting paid posters. Zeng et al. [28] investigated the behavior patterns of paid posters in online forums.

3 Detecting Organized Posters

3.1 Typical Organization Structure

To promote a campaign, the organizers of the campaign will typically employ three teams working for them: resource team, poster team and observation and evaluation team. The typical organization structure for paid posters is shown in Fig. 1. The organizers ask the resource team to prepare content of tweets for posting. The content can be not only text content, but also image, audio and even video. There are writers, graphic designers, video makers and so on in the resource team. Poster team is responsible for publishing the content manufactured by the resource team in popular websites like SINA Weibo. The observation and evaluation team is responsible for observing and evaluating the effect of the whole promoting activities and competitors' activities.

Fig. 1. Typical structure for the paid posters

The poster team mainly comes from two sources. First, some companies and organizations control large number of paid posters directly. These paid posters either controlled through open API of the platforms such as SINA Weibo Open Platform or employees in the company or organization. Second, some paid posters come from temporary recruitment. There are some platforms for hiring

part-time posters, such as Shuijunwang.com and 51shuijun.net. A company or organization can quickly employ a large number of paid posters from these platforms. The paid posters can be hired to attract public attention to their targets, enhance the strength of their viewpoints to something or even perturb public perspective. Some messages we see sometimes can not be trustworthy due to many rumors posted by them.

3.2 Framework for Detecting Organized Posters

Our framework for detecting paid posters is shown in Fig. 2 based on the individual and group characteristics using SVM model (IGCSVM). Given a user, we first study its individual statistical characteristics and group characteristics. The four individual characteristics and six group characteristics form a 10-dimensional vector. The features in the 10-dimensional vector are normalized to be between 0 and 1. Then we build a classification model from the training dataset to classify a user to be a paid poster or a legitimate user. A record in the training data is represented as the 10-dimensional vector and a class label (1 or -1). Class label 1 represents user u to be a paid poster and -1 represents it to be a legitimate user.

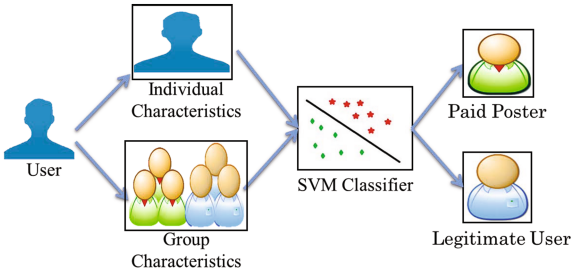


Fig. 2. Framework for detecting paid posters

Individual Statistical Characteristics. The four individual statistical characteristics are discussed in this section.

The Ratio of Friends to Followers. Some paid posters are not likely to be followed by normal users since they always do not post high quality contents. So they can not get many followers. The ratio of friends to followers (RFF) of a paid poster is always larger than normal users. We define the ratio of friends to followers P_{RFF} in Eq. 1.

$$P_{RFF} = \frac{N_{FR}}{N_{FR} + N_{FO}} \tag{1}$$

where N_{FR} is the number of friends of a user and N_{FO} is the number of followers of the user.

The Ratio of Tweets that Contain URLs to User's All tweets. Since the length of a tweet is not allowed to exceed 140 characters in microblogging websites like Twitter and SINA Weibo, there is always an URL in paid posters' tweets to promote a campaign. The ratio of tweets that contain URLs to user's all tweets (URL) for paid posters is probably higher than normal users. Equation 2 is defined to be the ratio of tweets that contain URLs to all tweets P_{URL}.

$$P_{URL} = \frac{N_{URL}}{N_{All}} \tag{2}$$

where N_{URL} is the number of tweets that contain URLs and N_{All} is the total number of tweets of a user.

The Ratio of Replied/Retweeted Tweets to User's All Tweets. Paid posters' tweets are less likely to be replied or retweeted than normal users' tweets. The first reason is that paid posters tend to post low quality tweets. The second one is that there are probably fewer normal users following them. Then the ratio of replied/retweeted tweets to user's all tweets (RRE) can be used distinguish paid posters and normal users. Equation 3 shows how to calculate the ratio of replied/retweeted tweets to user's all tweets P_{RRE}.

$$P_{RRE} = \frac{|TSet_{reply} \cup TSet_{retweet}|}{N_{All}} \tag{3}$$

where $TSet_{reply}$ and $TSet_{retweet}$ are the set of tweets that have been replied or retweeted. N_{All} is the total number of tweets for a user.

Influence. Ding et al. [10] compute a user's influence based on the multi-relational network. They perform multi random walks on the retweet, reply, reintroduce, and read networks which are constructed by the retweet, reply, reintroduce, and read relations between users. We implement their method on a multi-relational network that is constructed from the retweet and notify (@username) relations. There are more than 30 million users and a parallel distributed framework MapReduce[2] is used to compute the influence of users on a Hadoop[3] cluster which contains 32 nodes. The influence of a user (IN) $P_{IN}(0 \leq P_{IN} \leq 1)$ is defined to be a feature for detecting paid posters.

Group Characteristics. The six group characteristics are discussed in this section.

Original Tweet Posting. Paid posters tend to post copied tweets (sometimes changing few words) from the resource team which is described in Sect. 3.1 We call this feature "original tweet copying" (OTCopy). This observation has been widely studied in some existing researches [13,27] for detecting spam. To find the copied tweets, we first segment tweets to process Chinese words using ICTCLAS which is

[2] MapReduce: http://en.wikipedia.org/wiki/MapReduce.
[3] Apache Hadoop: http://en.wikipedia.org/wiki/Apache_Hadoop.

developed by Institute of Computing Technology, Chinese Academy of Sciences[4]. Then stopwords are removed and TF-IDF weighting schema is used to calculate weights of words. Finally we use vector space model (VSM) [21] to compute the similarity of two tweets. The threshold in our experiment is set to be 0.85, which is an empirical value, to determine whether two original posts (not a retweet) are the same. For a tweet $tweet_i$, we think it is copied from $tweet_j$ if the similarity between $tweet_i$ and $tweet_j$ is beyond the threshold and the posting time of $tweet_i$ is after $tweet_j$. We compared all tweets in our experiments to find copied tweets. Suppose a user u posts a total of N_{OT} tweets in a campaign, there are N_{OTCopy} tweets that are copied from others in a campaign. Then group characteristics "original tweets posting" P_{OT} for building classification model is obtained from the ratio of N_{OTCopy} and N_{OT} as shown in Eq. 4.

$$P_{OTCopy} = \frac{N_{OTCopy}}{N_{OT}}. \tag{4}$$

Retweeting. A retweet is a reposting of someone else's tweet. It is common to retweet its friends' tweets which can be seen in its timeline in SINA Weibo and add some comments on them. But for paid posters, they always retweet from someone who they do not follow and add the same comments that come from the resource team as other paid posters. Suppose a user u retweets a total of N_{RT} tweets, there are $N_{RTNonFriends}$ tweets that are retweeted from users who are not its friends, then the feature $P_{RTNonFriends}$ of group characteristic "retweeting without following (RTNonFriends)" for building classification model is obtained from the ratio of $N_{RTNonFriends}$ and N_{RT} as shown in Eq. 5.

$$P_{RTNonFriends} = \frac{N_{RTNonFriends}}{N_{RT}} \tag{5}$$

Suppose there are N_{RTCopy} tweets that have the same comments with others, then the feature "retweeting copy (RTCopy)" P_{RTCopy} for building classification model is obtained from the ratio of N_{RTCopy} and N_{RT} as shown in Eq. 6. The VSM model is used to measure if two comments are the same one like what has been done in measuring if two original tweets are the same ones.

$$P_{RTCopy} = \frac{N_{\mathrm{RTCopy}}}{N_{RT}} \tag{6}$$

Replying. Everyone can reply tweets in SINA Weibo. Like posting a new tweet, paid posters tend to get the comments from the resource team and they post the same comments (sometimes changing few words) on the target tweets. VSM model is also used to measure the similarity between two comments in a dataset. Paid posters are more likely to comment on users' tweets and the users are not their friends (non-friends). Given a user u who replies N_{RE} times in all tweets of a special campaign, there are $N_{RENonFriends}$ comments replied to non-friends'

[4] ICTCLAS: http://ictclas.org/index.html.

tweets, then the feature $P_{RENonFriends}$ of group characteristic "replying without following (RENonFriends)" for building classification model is obtained from the ratio of $N_{RENonFriends}$ and N_{RE} as shown in Eq. 7.

$$P_{RENonFriends} = \frac{N_{\text{RENonFriends}}}{N_{RE}} \tag{7}$$

If there are N_{RECopy} comments are the same as others, the feature P_{RECopy} of group characteristic "replying copy (RECopy)" is obtained from the ratio of N_{RECopy} and N_{RE} as shown in Eq. 8.

$$P_{RECopy} = \frac{N_{RECopy}}{N_{RE}} \tag{8}$$

Mentioning. Mentioning someone enables the mentioned user to receive a notification. It's also a convenient way for normal users to communicate with friends, but paid posters utilize the way to spread messages to the users they want. It's an usual way for paid posters to make others to see their tweets. This feature is also used to detect spammers in many studies [14,26,27]. If a user posts, retweets, replies the same tweet and mentions someone in its tweet, but the mentioned users are neither talked in the tweet nor followed by the poster, then it will be considered to be an abnormal action. Posting, retweeting and replying the same tweet has been studied in this section, we only consider the retweeting action with no comments but mentioning un-followed and un-related users in this paper. Given a user u who mentions un-followed and un-related users $N_{NoFollow}$ times in all N_{ME} tweets of a campaign and we call this feature" mentioning without following (NoFollow)", then the feature "mentioning without following (NoFollow)" P_{ME} can be obtained from the ratio of $N_{NoFollow}$ and N_{ME} as shown in Eq. 9.

$$P_{ME} = \frac{N_{NoFollow}}{N_{ME}}. \tag{9}$$

4 Experiments and Evaluation

4.1 Dataset

SINA Weibo[5], which is a microblogging website like Twitter, is one of the most popular websites in China with over 500 million registered users [1]. We collected public tweets via API in Sina Weibo. We obtained three datasets which are "Sina Campaign", "The Continent" and "Sangfor Tournament". The "Sina Campaign" dataset is conducted to promote a campaign in SINA Weibo. We collected all tweets about "Sina Campaign". To protect privacy, we do not show details in this dataset. We also collect two open public datasets "The Continent" and "Sangfor Tournament". We show the details about how we collected the two datasets. We extracted tweets that contain hashtag "#The Continent#"

[5] SINA Weibo: http://www.weibo.com/.

for dataset "The Continent". We collected 79,075 tweets from 72,064 users and 42,325 comments for the tweets between June 25 and July 25, 2014. Dataset for topic "Sangfor Tournament" was collected from tweets that contain keyword "Sangfor Tournament" from Jun 27 to Aug 27, 2014. There are 57,474 tweets from 16,364 users and 1,021 comments in the dataset. The follower/friend relationship and the most recent 200 tweets of all users in the three datasets were crawled.

Since it is hard to know who is exactly a paid poster or a legitimate user, to construct test datasets from topic 'The Continent" and "Sangfor Tournament", we randomly selected 450 users from each dataset and estimated them manually by three volunteers. They were asked to carefully check the content, the client, content of comments, retweeters of the top-100 posts of each user to evaluate whether a user was a paid poster or not. We also asked them to check other features like the user influence, the ratio of friends to followers, the ratio of replied/retweeted tweets to user's all tweets, the ratio of tweets that contain urls to user's all tweets and so on. For example, a user posts a tweet and the content of the tweet is the same as others (We set the number of persons to be 3 in our evaluation), and the client for posting the tweet is not coming from a sharing source like news website. Furthermore, the influence of the user, the ratio of friends to followers, the ratio of replied/retweeted tweets to user's all tweets are low, and the ratio of tweets that contain urls to user's all tweets is very high, then the user is probably a paid poster. If two or all of the three volunteers think the user is a paid posters, then it is. Otherwise, it is a legitimate user. There are 171 paid posters and 279 legitimate users in the "The Continent" dataset, comparing to 351 paid posters and 99 legitimate users in the "Sangfor Tournament" dataset.

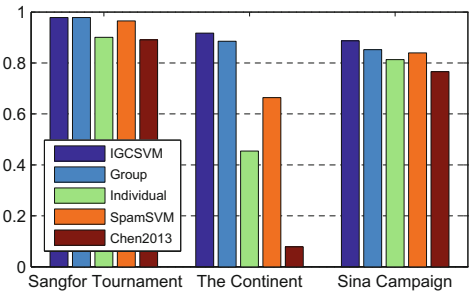
For dataset "Sina Campaign", we totally control the dataset and know who are the paid posters. We also randomly select 450 users like the datasets "The Continent" and "Sangfor Tournament". There are 294 paid posters and 156 legitimate accounts.

4.2 Experiments

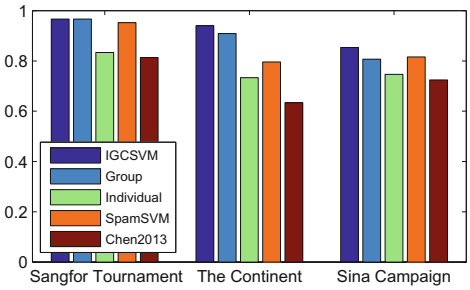
To evaluate the performance of our methods for detecting paid posters, we compare them with two baseline methods: SpamSVM method [4,18] and Chen2013 method [8]. 10-fold cross-validation is performed to analyze the performance of these methods in all experiments. Details of these methods are described below:

IGCSVM Method. Our IGCSVM method is based on both the individual statistical characteristics and group characteristics discussed in Sect. 3.2. Support Vector Machine (SVM) with a linear kernel was used to learn the classification model from the 10 features in Sect. 3.2. The values of the 10 features are computed by the equations in Sect. 3 like Eq. 1 and so on.

Individual method. Individual method is like the IGCSVM method, but it is only based on the four individual statistical characteristics of paid posters in Sect. 3.2.



(a) Average F1



(b) Average Accuracy

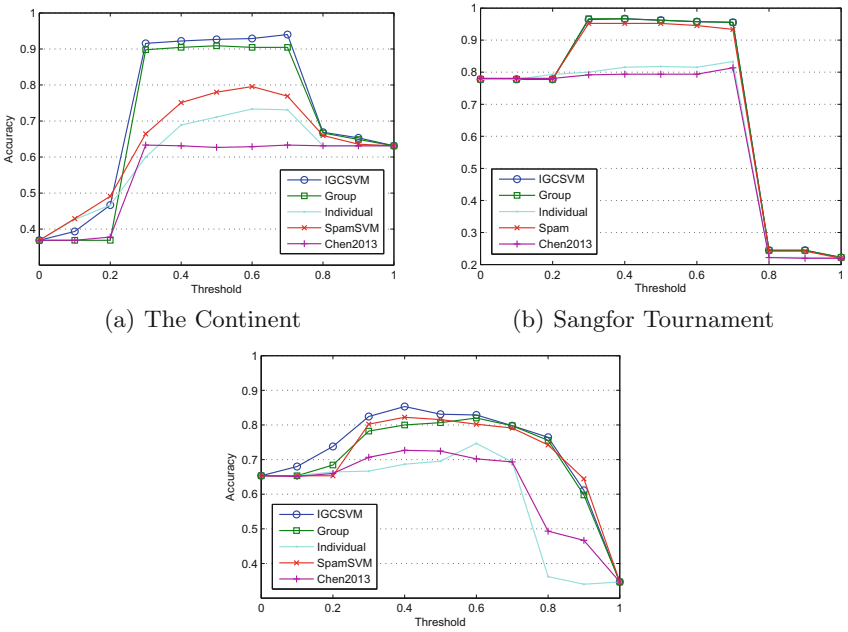
Fig. 3. Performance of the five methods

Group Method. Group method is like the IGCSVM method, but it is only based on the six group characteristics of paid posters in Sect. 3.2.

SpamSVM Method. Methods for detecting spammers can also be used to detect paid posters. Some researches [4,18] employ profile-based features and user's tweets to build an effective supervised learning model. A classifier is used to learn the model. And then the model is applied on unseen data to filter social spammers. In our experiments, profile-based features which are statistical features in Sect. 3.2 and semantic features which are original tweet copying and replying copy in Sect. 3.2 are employed.

Chen2013 Method. Chen et al. [8] proposed a method to detect paid posters using users' comments. Their method is based on users' comments rather than user's posts. The features they use in their method are ratio of replies, average interval time of posts, active days, the number of news reports and replying copy. LIBSVM [7] is also used in our experiments.

Support Vector Machine (SVM) with a linear kernel is used in all our experiments to learn classification models as it can get state of the art results [22]. SVM is a supervised learning model for classification and regression analysis. An open source implementation of SVM named LIBSVM [7] was used in all our experiments. LIBSVM is an integrated software for support vector classification and the main features of LIBSVM include different SVM formulations,



(a) The Continent (b) Sangfor Tournament

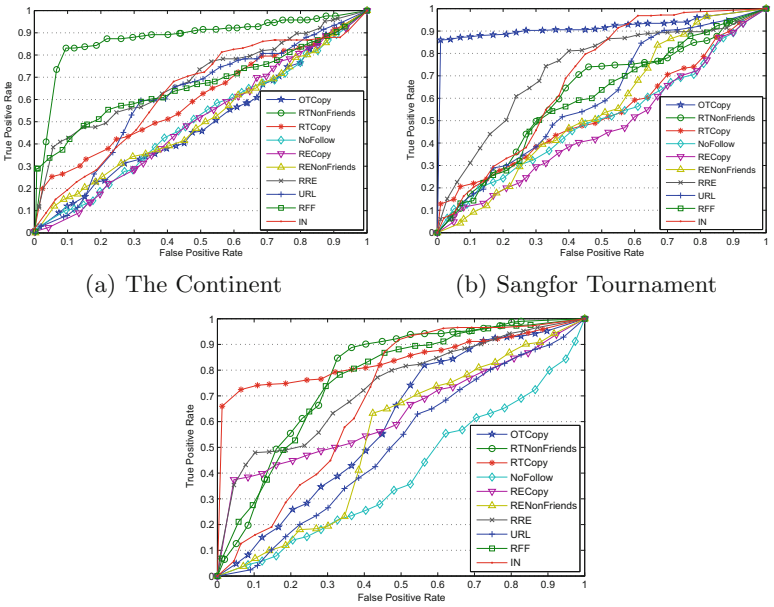
(c) Sina Campaign

Fig. 4. Accuracy comparison with the change of the threshold

efficient multi-class classification, cross validation for model selection, various kernels (including precomputed kernel matrix) and so on.

We compare the five methods in the datasets "The Continent", "Sangfor Tournament" and "Sina Campaign" with accuracy, and F1 score. Figure 3(a) and (b) show the performance results of the five methods in the three datasets. We can find that our ISCSVM method achieves the best performance on F1 score and accuracy in all the three datasets. It's significantly better than traditional spam detection method SpamSVM on F1 score and accuracy. The Group method is also better than traditional spam detection method SpamSVM and Individual method on F1 score and accuracy in all the three datasets. It shows that group features are more discriminative than traditional individual features for detecting spam in detecting paid posters. Chen2013 method is not good enough partly because there are only 1021 comments in the dataset "Sangfor Tournament".

We compare the accuracy of the five methods with the change of the threshold value which is used to distinguish ranges of values for detecting paid poster. If the value of the model predicting the probability of a user to be a paid poster is below the threshold, then it will considered to be a paid posters. Otherwise, it will be considered to be a legitimate user. The results on the three datasets are shown in Fig. 4. We can find that IGCSVM method gets the best performance when the threshold is between 0.3 and 0.7.



(a) The Continent (b) Sangfor Tournament

(c) Sina Campaign

Fig. 5. Features comparison

A Receiver Operating Characteristics (ROC) curve is constructed to measure the discrimination power of individual and group characteristics shown in Sect. 3. ROC curve is plotting true positive rate to false positive rate with the change of different threshold value. The four individual characteristics which are "RFF", "RRE", "URL" and "IN" and six group characteristics which are "OTCopy", "RTCopy", "RTNonFriends", "RECopy", "RENonFriends" and "NoFollow" are compared. Figure 5 shows the discrimination power of the ten features.

For the "The Continent" dataset shown in Fig. 5(a), we can find that "RTNonFriends" is the most discriminative feature in detecting paid posters. Features "NoFollow", "RECopy", "RENonFriends" and "OTCopy" are the least discriminative features. In the dataset "Sangfor Tournament" shown in Fig. 5(b), group feature "OTCopy" and individual feature "RRE" and "IN" are the most discriminative features in detecting paid posters. For the "Sina Campaign" dataset shown in Fig. 5(c), we can find that group feature "RTCopy", "RTNonFriends" and individual feature "RFF", "RRE" are the most discriminative feature in detecting paid posters. It shows that group features and individual features are both important to detect paid posters in dataset "Sina Campaign". It is the reason that our IGCSVM using both group and individual features gets better performance than Group method and Individual method which is based on only group or individual features.

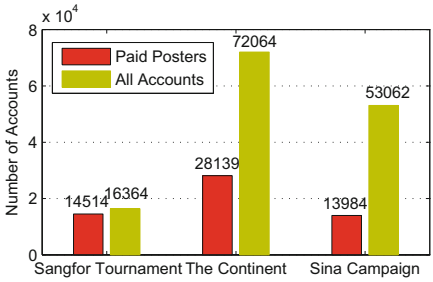


Fig. 6. Number of paid posters detected

We detect paid posters in the three datasets using IGCSVM method which gets the best accuracy and F1 score. The number of paid posters detected by IGCSVM method is shown in Fig. 6. IGCSVM method detects 14,514 paid posters in dataset "The Continent" which contains 16,364 users totally. It is 88.69 % of all users. It finds 28,139 paid posters in dataset "Sangfor Tournament", which is 39.05 % of all users. In "Sina Campaign" dataset, IGCSVM method detects 13,984 paid posters of totally 53,062 users, which is 26.35 % of all users.

5 Conclusion and Future Work

In this paper, we study a special type of online users named paid posters who are organized to post for purposes like advertising and so on in SINA Weibo. Our study is main related to online spam detection in social network. Our method utilizes the group characteristics of paid posters to detect them. Traditional individual statistics characteristics for detecting spam are also used to improve the performance. Our experimental results on the three datasets "Sangfor Tournament", "The Continent" and "Sina Campaign" show that group characteristics are also important in detecting paid posters comparing to traditional individual features. Our IGCSVM method which combines the two types of characteristics is effective in detecting paid posters and better than exiting approaches.

Our method in choosing features for detecting paid posters is empirical. It's better to learn effective features automatically to adapt to the change of paid posters. We will also try to improve the efficiency of our methods in future. For example, our methods based on the bag of words model have to compare all tweets in a dataset, it is not efficient enough. In future, we will try fingerprint based method and construct an index like B-tree to reduce the computational complexity.

Acknowledgments. This work was supported by 973 Program of China (Grant No. 2013CB329601, 2013CB329602, 2013CB329604), NSFC of China (Grant No. 60933005, 91124002), 863 Program of China (Grant No. 2012AA01A401, 2012AA01A402), National Key Technology RD Program of China (Grant No. 2012BAH38B04, 2012BAH38B06).

References

1. Sina weibo. http://en.wikipedia.org/wiki/sina_weibo, June 2014
2. Androutsopoulos, I., Koutsias, J., Chandrinos, K.V., Spyropoulos, C.D.: An experimental comparison of naive Bayesian and keyword-based anti-spam filtering with personal e-mail messages. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 160–167. ACM (2000)
3. Benevenuto, F., Duarte, F., Rodrigues, T., Almeida, V.A., Almeida, J.M., Ross, K.W.: Understanding video interactions in youtube. In: Proceedings of the 16th ACM international conference on Multimedia, pp. 761–764. ACM (2008)
4. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting spammers on twitter. In: Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS), vol. 6, p. 12 (2010)
5. Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Zhang, C., Ross, K.: Identifying video spammers in online social networks. In: Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web, pp. 45–52. ACM (2008)
6. Blanzieri, E., Bryl, A.: A survey of learning-based techniques of email spam filtering. Artif. Intell. Rev. **29**(1), 63–92 (2008)
7. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**(3), 27:1–27:27 (2011)
8. Chen, C., Wu, K., Srinivasan, V., Zhang, X.: Battling the internet water army: detection of hidden paid posters. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 116–120. ACM (2013)
9. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg? In: Proceedings of the 26th Annual Computer Security Applications Conference, pp. 21–30. ACM (2010)
10. Ding, Z., Jia, Y., Zhou, B., Han, Y.: Mining topical influencers based on the multi-relational network in micro-blogging sites. China Commun. **10**(1), 93–104 (2013)
11. Drucker, H., Wu, S., Vapnik, V.N.: Support vector machines for spam categorization. IEEE Trans. Neural Netw. **10**(5), 1048–1054 (1999)
12. Fetterly, D., Manasse, M., Najork, M.: Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In: Proceedings of the 7th International Workshop on the Web and Databases: Colocated with ACM SIGMOD/PODS 2004, pp. 1–6. ACM (2004)
13. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, pp. 35–47. ACM (2010)
14. Grier, C., Thomas, K., Paxson, V., Zhang, M.: @ spam: the underground on 140 characters or less. In: Proceedings of the 17th ACM Conference on Computer and Communications Security, pp. 27–37. ACM (2010)
15. Gyöngyi, Z., Garcia-Molina, H., Pedersen, J.: Combating web spam with trustrank. In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDB Endowment, vol. 30, pp. 576–587 (2004)
16. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230. ACM (2008)

17. Kolari, P., Java, A., Finin, T., Oates, T., Joshi, A.: Detecting spam blogs: a machine learning approach. In: Proceedings of the National Conference on Artificial Intelligence, vol. 21, pp. 1351. AAAI Press, Menlo Park, CA (1999), MIT Press, Cambridge, London, MA (2006)
18. Lee, K., Caverlee, J., Webb, S.: Uncovering social spammers: social honeypots+ machine learning. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 435–442. ACM (2010)
19. McCord, M., Chuah, M.: Spam detection on twitter using traditional classifiers. In: Alcaraz Calero, J.M., Yang, L.T., Mármol, F.G., Villalba, L.J.G., Li, A.X., Wang, Y. (eds.) ATC 2011. LNCS, vol. 6906, pp. 175–186. Springer, Heidelberg (2011)
20. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 309–319. Association for Computational Linguistics (2011)
21. Salton, G., Wong, A., Yang, C.-S.: A vector space model for automatic indexing. Commun. ACM **18**(11), 613–620 (1975)
22. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)
23. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, pp. 243–258. ACM (2011)
24. Thomason, A.: Blog spam: a review. In: CEAS (2007)
25. Wang, K., Xiao, Y., Xiao, Z.: Detection of internet water army in social network. In: 2014 International Conference on Computer, Communications and Information Technology (CCIT 2014). Atlantis Press (2014)
26. Zhang, Y., Ruan, X., Wang, H., Wang, H.: What scale of audience a campaign can reach in what price. In: 2014 IEEE International Conference on Computer Communications (InfoCOM 2014) (2014)
27. Yang, C., Harkreader, R., Zhang, J., Shin, S., Gu,G.: 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, pp. 71–80. ACM (2012)
28. Zeng, K., Wang, X., Zhang, Q., Zhang, X., Wang, F.-Y.: Behavior modeling of internet water army in online forums. World Congr. **19**, 9858–9863 (2014)