Finding the Hidden Hands: A Case Study of Detecting Organized Posters and Promoters in SINA Weibo

WANG Xiang[1], ZHANG Zhilin[2], YU Xiang[1], JIA Yan[1,3], ZHOU Bin[1,3], LI Shasha[1]

[1] School of Computer, National University of Defense Technology, Changsha, Hunan 410073, China

[2] School of Computing Science, Simon Fraser University, Burnaby, Canada

[3] State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha, 410073, China

Abstract: With the development of online social networks, a special group of online users named organized posters (or Internet water army, Internet paid posters in some literatures) have flooded the social network communities. They are organized in groups to post with specific purposes and sometimes even confuse or mislead normal users. In this paper, we study the individual and group characteristics of organized posters. A classifier is constructed based on the individual and group characteristics to detect them. Extensive experimental results on three real datasets demonstrate that our method based on individual and group characteristics using SVM model (IGCSVM) is effective in detecting organized posters and better than existing methods. We take a first look at finding the promoters based on the detected organized posters of our IGCSVM method. Our experiments show that it is effective in detecting promoters.

Keywords: organized posters; internet water army; online paid posters; promoter; micro-blogging

I. INTRODUCTION

Today, various social networks like Twitter, Facebook, and Flickr are becoming increasingly popular information sources for billions of people. Due to the ease of forwarding messages, information can be widely disseminated instantly to interested people via one's social networks. A group of users, called "organized posters" [1] in this paper, are recruited to engage with other normal users for increasing awareness of their tweets. Different from normal users, organized posters are always paid to post, reply and retweet some specific messages. Suffering from the flooding of these deceptive messages into microblogging platforms, normal users can get incorrect even contrary impression of some certain affairs and events. Normal users can be confused or misled due to large number of non-objective messages from them. In the worse cases, more serious consequences, like film box office changing, stock market disruption or widespread panic, are raised by the information. For example, a famous Chinese film-maker and screenwriter named Lu Chuan announced that his film "The Last Suppe" was attacked by the organized posters in social network . There were a large number of slanderous and bad comments for the film in social network. Large number of normal users read the negative comments and did not watch the film in cinema, so the film lost its box office. To reduce the negative effect, it's crucial for us to detect organized posters and

analyze their group characteristics.

Organized posters are different from traditional spammers. First, typical organized posters are well organized and can bring great harm to some persons, companies and organizations. Some organized poster groups go far away than posting spam messages, the behaviors of them sometimes are illegal and disrupt the normal life of Internet users. In Internet, organized posters are somewhat like organized criminal groups in our real life that we have to fight with. Second, organized posters are either controlled by a program through platform API or human beings. They are different from Twitter bot [2] 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. Third, organized posters are more covert than spammers. They are normal users at ordinary times, but they become organized posters when they try to promote a campaign. Even Some famous users with high influence can be paid to be organized posters temporarily when they are needed in a promoting campaign. Opinion spam is a kind of organized posters [3] [4], but existing researches focused on detect them in electronic-commerce websites like Amason and hotel booking website TripAdvisor, rather than social network platforms like microblogging websites.

There are many studies about detecting spammers and analysis their characteristics [5] [6] [7]. They detect spammers by clustering URLs [8] [9], similarity of microblogging text [8] or users' mention action [7]. They are mainly concerned on individual characteristics like user profiles for detecting spam. But group characteristics are important for detecting organized posters. For example, given a business promotion campaign for promoting a website, a large number of organized posters are paid to retweet an advertising tweet to their communities and 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 organized posters.

In this paper, we study several useful group characteristics for detecting organized posters. Some individual characteristics used in traditional spam detecting methods are also utilized in our method. Our method combines both the individual characteristics and group characteristics to detect organized posters. Experiments on three real datasets show that the proposed method is effective in detecting organized posters.

We also try to find the promoters (the hidden hands) in a campaign based on the detected organized posters. We define promoters to be the authors of the source of the promoting tweets in a campaign. We detect promoters based on the detected organized posters and the propagation graphs of the promoting tweets.

Our main contributions can be summarized as follows: (1) We describe the typical organization structure of organized posters in a promoting campaign. We propose a SVM based method named "IGCSVM" using both user's individual and group characteristics for detecting organized posters. We find that group characteristics are more important than traditional individual features in detecting organized posters. (2) We also take a first look at detecting promoters in a campaign based on results of detected organized posters. (3) Extensive experiments have been done on three real datasets crawled from SINA Weibo. Experimental results show that our IGCSVM method is more effective than existing methods in detecting organized posters and the features we choose for detecting organized posters are effective. Experimental results also show that our method is effective in detecting promoters.

The rest of this paper is organized as follows: Section 2 discusses some important related works. Section 3 introduces our method for detecting organized posters. Experimental results are shown in Section 4. Section 5 discusses how to find promoters of a campaign. Finally, conclusion and future work are provided in section 6.

In this paper, The authors study the individual and group characteristics of organized posters. A classifier is constructed based on the individual and group characteristics to detect them.

II. Related Works

Spammers have been appearing in a lot of applications, such as blogs [11] [12], email [13]

[14], Web search engine [15] and videos [16] [17]. And there are a large amount of methods which have been proposed to detect them [18] [19]. Zhang et al. [7] 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 analyze 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. Yanget al. [5] 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. [8] proposed a method to detect malicious users and posts based on URL and text clustering. They also analyze the characteristics of the malicious users and posts. Thomas et al. [20] 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. [21] 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. [6] studied spam on Twitter and found that click through rate of spam URLs was much lower than email. The analysis also showed that 84% spam users are organized by a few controllers. M. McCord and M. Chuah [22] 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. [2] build a classifier to determine an account to be a human, bot or cyborg.

There are also some researches about organized posters. Opinion spam is a kind of organized poster. Jindal and Liu [3] find that opinion spam is widespread and in electronic commerce websites. They train their models using features like review text, reviewer and product to detect duplicate opinions in Amazon. Ott et al. [4] proposed n-gram based text categorization to detect deceptive opinion spam in hotel booking website TripAdvisor. Chen et al. [1] investigated the behavioral pattern of organized posters and designed a detection mechanism to identify potential organized 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 organized posters in this paper. Wang et al. [23] studied five features for detecting organized posters. Zeng et al. [24] investigated the behavior patterns of organized posters in online forums.

III. 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 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. They also have to analyze competitors' activities. The organization structure for promoting a campaign is shown in Figure 1.

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

of organized posters comes 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 organized posters from these platforms. The organized posters are hired to attract public attention to their targets, enhance the strength of their viewpoints or perturb public perspective. Many messages we see sometimes can not be trustworthy due to many rumors posted by them.

3.2 Problem statement

In microblogging website like SINA Weibo, there are k users $U=\{u_1,u_2,\cdots,u_k\}$ in a campaign or a topic. Each user u_i posts, retweets and replies a number of tweets. They also follow some users and are followed by others. The organized poster detection problem is to estimate whether u_i is an organized poster through a classification model c. A classification model $c: u_i \rightarrow \{Organized\,\text{Poster}, \text{LegitimateUser}\}$ predicts whether u_i is an organized poster. We need to find a set of features F from the users in the campaign to train classification model c. We choose two types of features for detecting organized posters: the individual statistical characteristics and the group characteristics.

3.3 Framework for detecting organized posters

3.3.1 Individual statistical characteristics

The four individual statistical characteristics are discussed in this section.

The Ratio of Friends to Followers. Some organized 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 an organized poster is probably larger than normal users. We define the ratio of friends to followers P_{RFF} as Equation 1,

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

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

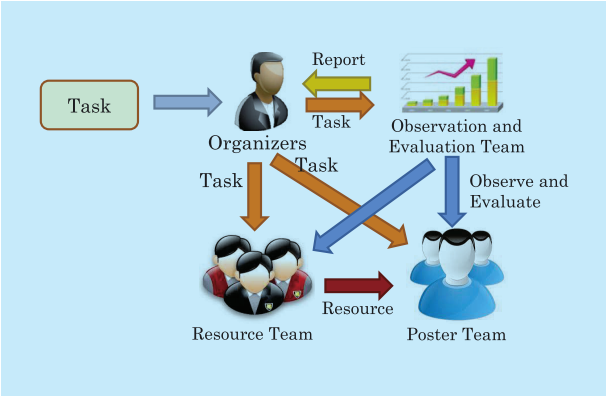


Fig.1 *Typical structure for the organized posters*

The Ratio of Tweets that Contain URLs to User's All tweets. There is always an URL in organized posters' tweets to promote a campaign, since the length of a tweet is not allowed to exceed 140 characters. The ratio of tweets that contain URLs to user's all tweets (URL) for organized posters is probably higher than normal users. Equation 2 is defined to compute 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. Organized posters' tweets are less likely to be replied or retweeted comparing to normal users' tweets. The first reason is that organized 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 organized posters and normal users. Equation 3 shows how to calculate the ratio of replied/retweeted tweets to user's all tweets P_{RRE}.

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

where $TSet_{reply}$ and $TSet_{retweeky}$ 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 is used to compute the influence of users on a Hadoop cluster which contains 32 nodes. The influence of a user (IN) P_{IN} $(0 \le P_{IN} \le 1)$ is defined to be a feature for detecting organized posters.

3.3.2 Group characteristics

The six group characteristics are discussed in this section.

Original Tweet Posting. Organized posters tend to post copied tweets (sometimes changing few words) from the resource team which is described in Section 3.1. We call this feature "original tweet copying" (OTCopy). This observation has been widely studied in some existing researches [8] [5] for detecting spammers. To find the copied tweets, we first segment tweets to process Chinese words using ICTCLAS which is developed by Institute of Computing Technology, Chinese Academy of Sciences . Then stopwords are removed and TF-IDF weighting schema is used to calculate weights of words. Finally we use vector space model (VSM) [25] to compute the similarity of two tweets. The threshold in our experiment is set to be 0.85, which is a empirical value, to determine whether two original posts (not a retweet post) 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 groups of copying 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 Equation 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 organized posters, they always retweet from someone who they do not follow and add the same comments that come from the resource team as other organized 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 the ratio of $N_{RTNonFriends}$ and N_{RT} as shown in Equation 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 Equation 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, organized 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. Organized 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' tweets, then the feature $P_{RENonFriends}$ of group characteristic "replying without following (RENonFriends)" for building classification model is obtained from the the ratio of $N_{RENonFriends}$ and N_{RE} as shown in Equation 7.

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

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 the ratio of N_{RECopy} and N_{RE} as shown in Equation 8.

$$P_{RECopy} = \frac{N_{RECopy}}{N_{RE}} \quad (8)$$

Mentioning. Mentioning someone enables the mentioned user to receive a notification. It's an usual way for organized posters to make others to see their tweets. It's a convenient way for normal users to communicate with friends, but organized posters utilize the way to spread messages to the users they want. This feature is also used to detect spammers in many studies [7] [6] [5]. If a user posts, retweets, replies the same tweet with others except the mentioned users and the mentioned users are neither talked in the tweet nor followed by the poster, 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 the ratio of $N_{NoFollow}$ and N_{ME} as shown in Equation 9.

$$P_{ME} = \frac{N_{NoFollow}}{N_{ME}} \quad (9)$$

3.3.3 Framework for detecting organized posters

The framework for detecting organized posters is shown in Figure 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 four individual characteristics are the ratio of friends to followers (RFF), the ratio of replied/retweeted tweets to user's all tweets (RRE), the ratio of tweets that contain URLs (URL) to user's all tweets and influence of the user (IN). The six group features for users to post in groups are "original tweets copy (OTCopy)", "retweeting copy (RTCopy)", "retweeting without following (RTNonFriends)", "replying copy (RECopy)", "replying without following (RENonFriends)" and "mentioning without following (NoFollow)". Table 1 shows the 10 characteristics and their explanation.

Table I *The 10 characteristics*

Characteristics	Explanation
RFF	RFF is probably larger than normal users
RRE	Organized posters' tweets are less likely to be replied or retweeted
URL	More promoting tweets (containing URLs) than normal users
IN	The influence of organized posters is probably lower than normal users
OTCopy	Organized posters tend to post copied tweets
RTCopy	Retweeting a tweet with copied comments
RTNonFriends	Retweeting from someone who they do not follow
RECopy	Replying a tweet with copied comments
RENonFriends	Replying a tweet from someone who they do not follow
NoFollow	Mentioning someone who is neither talked in the tweet nor followed by the poster

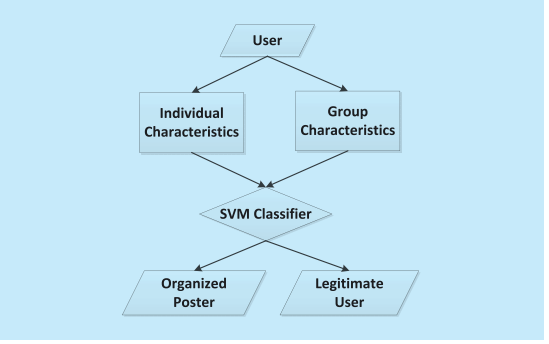


Fig.2 *Framework for detecting organized posters*

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 an organized poster or 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 an organized poster and -1 represents it to be a legitimate user. The individual features and group features are respectively discussed in Section 3.3.1 and 3.3.2.

IV. 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[6]. We collected public tweets via API in Sina

Weibo. We obtained three datasets which are "Sina Campaign", "The Continent" and "Sangfor Tournament". "Sina Campaign" 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#" 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.

The three campaigns of "The Continent", "Sangfor Tournament" and "Sina Campaign" fit the characteristics of promoting campaigns. For example, they all direct users to marketing URLs and mention a lot of users. The purpose of the three campaigns is to make the public known about the topics and join in them. 9,618 users of all 79,075 users, which are 12.16% of all users, are blocked by SINA Weibo platform in the topic "The Continent" when we check them on Oct. 22, 2014. There are 8,153 users in topic "Sangfor Tournament" blocked by SINA Weibo platform and they are 49.82% of all users. 3,209 users of all 53,062 users in the topic "Sina Campaign" are blocked by SINA Weibo platform while we check it on June 19, 2014. In other words, more than 6.05% of all users are blocked. The blocked users show that the campaign employs organized posters for the promoting goals. Figure 3 shows the number of blocked users in the three datasets.

Since it is hard to know who is exactly an organized poster or 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 an organized 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 an organized poster. If two or all of the three volunteers think the user is an organized posters, then it is. Otherwise, it is a legitimate user. There are 171 organized posters and 279 legitimate users in the "The Continent" dataset, comparing to 351 organized posters and 99 legitimate users in the "Sangfor Tournament" dataset.

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

4.2 Experiments

To evaluate the performance of our methods for detecting organized posters, we compare them with two baseline methods: SpamSVM method [21] [26] and Chen2013 method [1]. 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 method based on individual and group characteristics of SVM (IGCSVM) is based on both the individual statistical characteristics and group characteristics discussed in Section 3.3. Support Vector Machine (SVM) with a linear kernel was used to learn the classification model from the 10 features in Section 3.3. The values of the 10 features are

computed by the equations in Section 3 like Equation 1 and so on.

Individual method. Individual method is like the IGCSVM method, but it is only based on the individual statistical characteristics of organized posters in Section 3.3.1.

Group Method. Group method is like the IGCSVM method, but it is only based on the group characteristics of organized posters in Section 3.3.2.

SpamSVM Method. Methods for detecting spammers can also be used to detect organized posters. Some papers [21] [26] 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 Section 3.3.1 and semantic features which are original tweet copying and replying copy in Section 3.3.2 are employed.

Chen2013 Method. Chen et al. [1] proposed a method to detect organized 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 [27] is also used in our experiments.

Support Vector Machine (SVM) with a linear kernel was used in all our experiments to learn classification models as it can get state of the art results [28]. SVM is a supervised learning model for classification and regression analysis. An open source implementation of SVM named LIBSVM [27] 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, efficient multi-class classification, cross validation for model selection, Various kernels (including precomputed kernel matrix) and so on.

We compare the five methods in dataset "The Continent", "Sangfor Tournament" and "Sina Campaign" with accuracy, false positive rate (FPR) and F1 measure. Table 2, 3 and 4 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 in all the three datasets. The Group method is also better than traditional spam detection method SpamSVM on F1 Score and accuracy in all the three datasets. It shows that group features are more important than traditional individual features for detecting spam in detecting organized posters in all the three datasets. In the "Sangfor Tournament" dataset, we can find that the IGCSVM method with all

Table II *Performance results of the "Sangfor Tournament" dataset*

Method	FPR	F1 Score	Accuracy
IGCSVM	1.0%	0.9782	96.67%
Group	1.0%	0.9782	96.67%
Individual	65%	0.9007	83.33%
SpamSVM	0.0%	0.9653	95.23%
Chen2013	75.76%	0.8909	81.37%

Table III *Performance results of the "The Continent" dataset*

Method	FPR	F1 Score	Accuracy
IGCSVM	3.87%	0.9174	94%
Group	11.62%	0.8852	90.89%
Individual	1.41%	0.4545	73.33%
SpamSVM	5.99%	0.6642	79.56%
Chen2013	2.11%	0.0782	63.33%

Table IV *Performance results of the "Sina Campaign" dataset*

Method	FPR	F1 Score	Accuracy
IGCSVM	19.87%	0.8870	85.33%
Group	28.20%	0.8523	80.67%
Individual	43.59%	0.8131	74.66%
SpamSVM	3.85%	0.8395	81.56%
Chen2013	21.15%	0.7660	72.44%

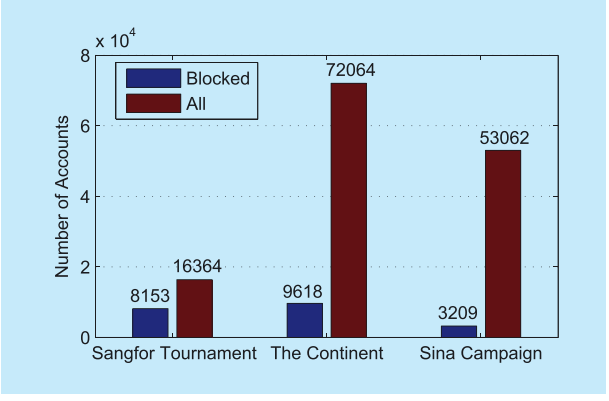
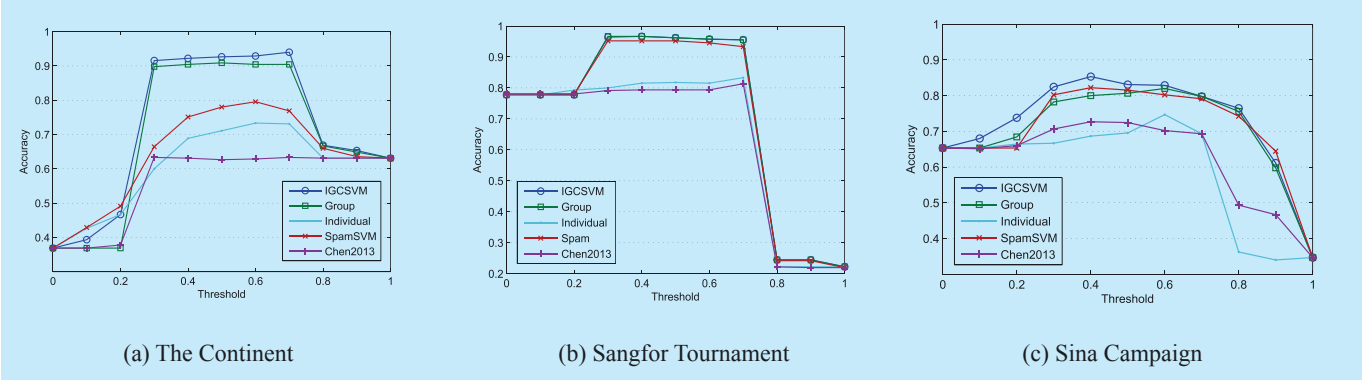
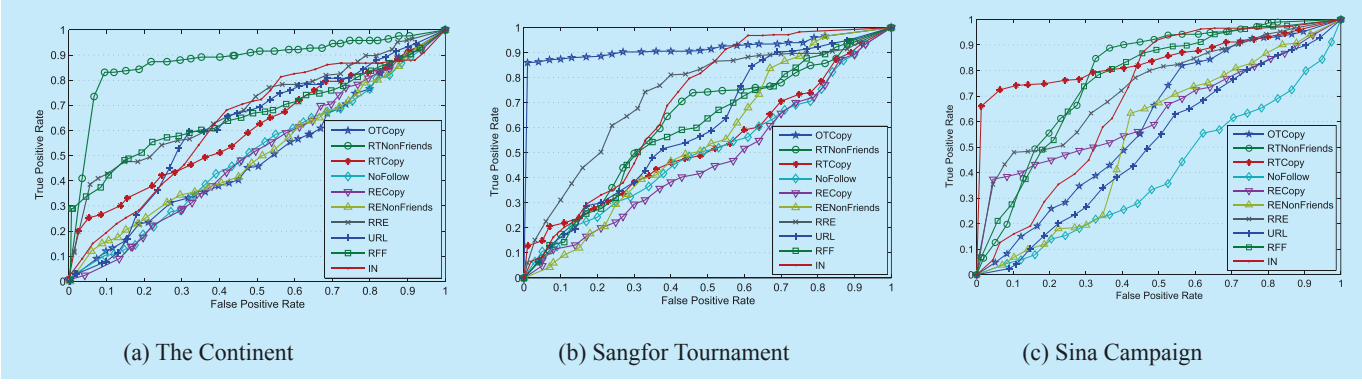


Fig.3 *Blocked-three datasets*



(a) The Continent (b) Sangfor Tournament (c) Sina Campaign

Fig.4 *Accuracy comparison*



(a) The Continent (b) Sangfor Tournament (c) Sina Campaign

Fig.5 *Features comparison*

features and Group method with group features get the best F1 score and accuracy at the same time. The Individual method is worse than the Group method on false positive rate, F1 score and accuracy. It shows that the group characteristics are more discriminative than the individual statistical characteristics. Chen2013 method is the worst one partly because there are very few comments in dataset "Sangfor Tournament". There are only 1021 comments comparing to 57,474 tweets in dataset "Sangfor Tournament". In dataset "The Continent", the IGCSVM method is significantly better than traditional spam detection method SpamSVM and Group method since it combines the individual statistical and group characteristics. The Individual method gets the best false positive rate. Chen2013 method which only based on comments gets worst F1 score and accuracy partly since there are only 42,325 comments which is only half of the number of tweets in the "The Continent" dataset. In dataset "Sina Campaign", the IGCSVM method, which combines the individual statistical and group characteristics, is significantly better than Group method and SpamSVM method.

We compare the accuracy of the five methods with the change of threshold value which is used to distinguish ranges of values for detecting organized poster. The results on the three datasets are shown in Figure 4. We can find that IGCSVM method gets the best performance when the threshold is between 0.3 and 0.7.

A Receiver Operating Characteristics (ROC) curve is constructed to measure the discrimination power of individual and group characteristics shown in Section 3. ROC curve is plotting true positive rate to false positive rate with the change of different threshold value. There are 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

Figure 5(a), we can find that "RTNonFriends" is the most discriminative feature in detecting organized posters. Features "NoFollow", "RECopy", "RENonFriends" and "OTCopy" are the least discriminative features. In dataset "Sangfor Tournament" shown in Figure 5(b), group feature "OTCopy" and individual feature "RRE" and "IN" are the most discriminative features in detecting organized posters. For the "Sina Campaign" dataset shown in Figure 5(c), we can find that group feature "RTCopy", "RTNonFriends" and individual feature "RFF", "RRE" are the most discriminative feature in detecting organized posters. It shows that group features and individual features are both important to detect organized 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.

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

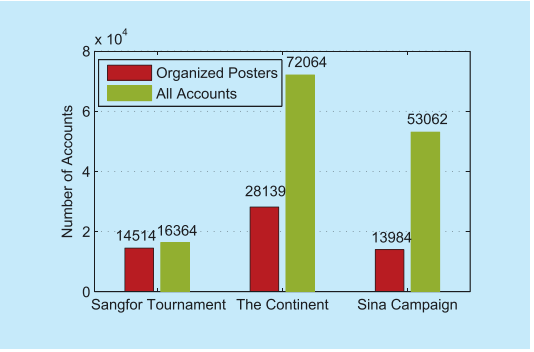


Fig.6 *Number of organized posters detected*

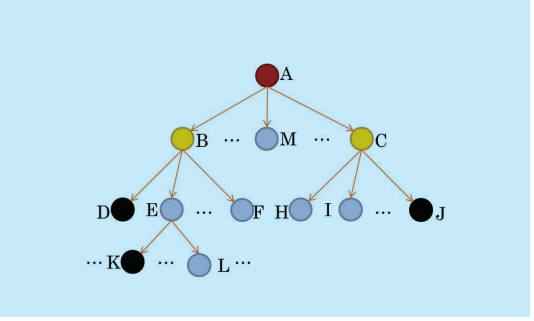


Fig.7 *The typical propagation graph of a tweet. Red node A is the source of the tweets. The yellow nodes like B and C are the copied tweets. The blue ones like M are tweets reposted from its father. The black nodes like D are the replied comments of tweets*

V. Finding the Promoters in a Campaign

In this section, we find the hidden hands or the promoters of a campaign. Promoters in this paper are the authors of the source of the promoting tweets in a campaign. In real world, it is interesting and useful to find the promoters

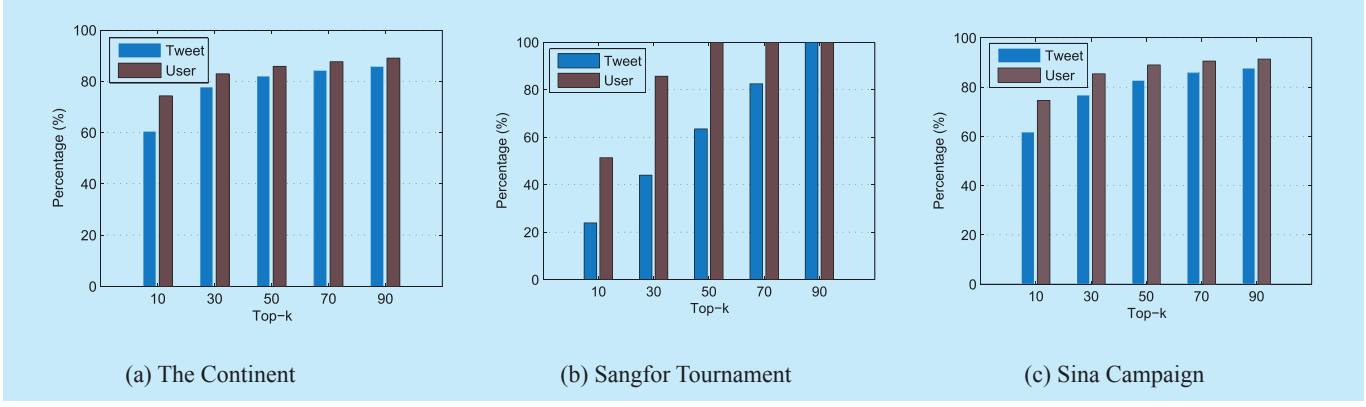


Fig.8 *Features comparison*

to know who are the promoters promoting the campaign. Although the promoters detected in this paper may be not the organizer or constitutors of the campaign in the real world, it still provides a key to find them, since promoters are the source of the promoting tweets. Sometimes the organizer or constitutors are among the list of promoters. We propose a simple but effective method for detecting promoters in a topic.

Our method is based on the detected organized posters by IGCSVM method in Section 4. Figure 6 shows the number of organized posters in the three datasets. The accuracies of IGCSVM method in the three datasets are all above 85%, so it is reasonable to reuse the detected results of our IGCSVM method.

Our hypothesis is that if many organized posters in a topic try to promote a tweet, then the author of the tweet is probably the promoter of the topic. We study three kinds of behavioral features in Microblogging website like Twitter and SINA Weibo for a user to promote a tweet. The three kinds of features are "original tweet copying", "retweeting" and "replying".

We construct propagation graphs to find the source of promoting tweets posted by organized posters with our IGCSVM method. A typical propagation graph of tweets constructed from "original tweet copying", "retweeting" and "replying" is shown in Figure 7. If a tweet j is copied or retweeted from tweet i by an organized poster, or j is the replied comments for tweet i by an organized poster, then there is an edge from i to j. We use vector space model (VSM) to measure if a tweet is copied from others like what we have done in computing if two original tweets are the same one in Section 3.3.2. Let there are a number of same tweets found by the VSM model, then the tweet t, whose publication time is the earliest, is the source of all the copied tweets. There are edges from t to all copied tweets.

A reverse depth first search method is used to find the number of tweets in a propagation graph. If there are N tweets in the propagation graph, then the number of tweets posted by organized posters to promote source tweet A $N_{TWeek(i)}$ can be calculated as Equation 10.

$$N_{Tweet}(\mathrm{i}) = \begin{cases} N-1, & \text{i is the source tweet,} \\ 0, & others \end{cases} \quad (10)$$

Figure 8 shows the percentage of organized posters who participating in the top-k source tweets and top-k users' source tweets in dataset "Sangfor Tournament", "The Continent" and "Sina Campaign". We detected 14,514 organized posters using our IGCSVM method in "Sangfor Tournament" dataset. There are totally 33973 times that they participate in others' tweets in dataset "The Continent". There are 20558 times for organized poster to participate in top-10 tweets, which is 60.51% of all times. For the top-10 users, there are 74.36% of total times that organized poster participate in. In the top-90 tweets, there are 85.77% of all times that organized poster participate in, comparing to 89.06% of total times in the top-90 users. It shows that most organized posters participate in few users' tweets. In dataset "Sangfor Tournament", there are totally 55424 times that they participate in others' tweets. In the top-10 tweets and users, there are 23.87% and 51.34% of all times that organized posters participate. But for the top-90 tweets and top-50 users, there are 99.99% and 99.98% of all times that organized posters participate. In dataset "Sina Campaign", there are totally 48,557 times that they participate in others' tweets. In the top-10 tweets and users, there are 61.76% and 74.57% of all times that organized posters participate. But for the top-90 tweets and top-90 users, there are 87.66% and 91.37% of all times that organized posters participate. The results in the three datasets validate that most organized posters actually participate in few users' tweets. In other words, the few users are probably promoters. In the three datasets, promoters are among the top-90 users in "The Continent" dataset, top-50 users in "Sangfor Tournament" dataset and top-90 users in "Sina Campaign", because over 89% of total times that organized posters participate in these few users' tweets in all the three datasets. So promoters are among a small number of users in all the three datasets.

VI. Conclusions and Future Work

In this paper, we study a special type of online users named organized posters who are organized

to post for purposes like advertising and so on in SINA Weibo. Our study is main related to online spammer detection in social network. Our method utilizes the group characteristics of organized posters to detect them. Traditional individual statistics characteristics for detecting spammers are also used to improve the performance. Our experimental results on three datasets “Sangfor Tournament", “The Continent” and “Sina Campaign” show that group characteristics are discriminative features in detecting organized posters. Our IGCSVM method is very effective in detecting organized posters and better than exiting approaches. Furthermore, we take a first look at finding the promoters in a campaign. Our method for detecting promoters is based on the organized posters detected by our IGCSVM method. Our experimental results show that most organized posters actually participate in very few users' tweets. Promoters are among a small number of users.

Our method in choosing features for detecting organized posters is empirical. It's better to learn effective features automatically to adapt to the change of organized 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 has to compare all tweets in a campaign, it is not effective enough. In future, we will try fingerprint based method and construct an index like B-tree to reduce the computational complexity.

Acknowledgements

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] Cheng Chen, KuiWu, VenkateshSrinivasan, and Xudong Zhang. 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, pages 116–120. ACM, 2013.

[2] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In Proceedings of the 26th annual computer security applications conference, pages 21–30. ACM, 2010.

[3] Nitin Jindal and Bing Liu. Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining, pages 219–230. ACM, 2008.

[4] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 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-Volume 1, pages 309–319. Association for Computational Linguistics, 2011.

[5] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei 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, pages 71–80. ACM, 2012.

[6] Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. @ spam: the underground on 140 characters or less. In Proceedings of the 17th ACM conference on Computer and communications security, pages 27– 37. ACM, 2010.

[7] Yubao Zhang Xin Ruan Haining Wang Hui Wang. What scale of audience a campaign can reach in what price. In 2014 IEEE International Conference on Computer Communications (InfoCOM'14), 2014.

[8] Hongyu Gao, Jun Hu, Christo Wilson, Zhichun Li, Yan Chen, and Ben Y Zhao. Detecting and characterizing social spam campaigns. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, pages 35–47. ACM, 2010.

[9] Kurt Thomas, Chris Grier, Justin Ma, Vern Paxson, and Dawn Song. Design and evaluation of a real-time url spam filtering service. In Security and Privacy (SP), 2011 IEEE Symposium on, pages 447–462. IEEE, 2011.

[10] Zhaoyun DING, Yan JIA, Bin Zhou, and Yi HAN. Mining topical influencers based on the multi-relational network in micro-blogging sites. China Communications, 10(1):93–104, 2013.

[11] Adam Thomason. Blog spam: A review. In CEAS, 2007.

[12] Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. Detecting spam blogs: A machine learning approach. In PROCEEDINGS OFTHENATIONALCONFERENCEONARTIFICIALINTELLIGENCE, volume21, page1351.MenloPark, CA;Cambridge, MA;London; AAAI Press; MIT Press; 1999, 2006.

[13] Enrico Blanzieri and Anton Bryl. A survey of

learning-based techniques of email spam filtering. Artificial Intelligence Review, 29(1):63–92, 2008.

[14] Ion Androutsopoulos, John Koutsias, Konstantinos V Chandrinos, and Constantine D Spyropoulos. 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, pages 160–167. ACM, 2000.

[15] Dennis Fetterly, Mark Manasse, and Marc Najork. 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, pages 1–6. ACM, 2004.

[16] Fabricio Benevenuto, Fernando Duarte, Tiago Rodrigues, Virgilio AF Almeida, Jussara M Almeida, and Keith W Ross. Understanding video interactions in youtube. In Proceedings of the 16th ACM international conference on Multimedia, pages 761–764. ACM, 2008.

[17] Fabricio Benevenuto, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, Chao Zhang, and Keith Ross. Identifying video spammers in online social networks. In Proceedings of the 4th international workshop on Adversarial information retrieval on the web, pages 45–52. ACM, 2008.

[18] Harris Drucker, S Wu, and Vladimir N Vapnik. Support vector machines for spam categorization. Neural Networks, IEEE Transactions on, 10(5):1048–1054, 1999.

[19] Zoltán Gy?ngyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, pages 576–587. VLDB Endowment, 2004.

[20] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, pages 243–258. ACM, 2011.

[21] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: social honeypots+ machine learning. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pages 435–442. ACM, 2010.

[22] M McCord and M Chuah. Spam detection on twitter using traditional classifiers. In Autonomic and trusted computing, pages 175–186. Springer, 2011.

[23] Kun Wang, Yang Xiao, and Zhen Xiao. Detection of internet water army in social network. In 2014 International Conference on Computer, Communications and Information Technology (CCIT 2014). Atlantis Press, 2014.

[24] Ke Zeng, Xiao Wang, Qingpeng Zhang, Xinzhan Zhang, and Fei-Yue Wang. Behavior modeling of internet water army in online forums. In World Congress, volume 19, pages 9858–9863, 2014.

[25] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620, 1975.

[26] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. Detecting spammers on twitter. In Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), volume 6, page 12, 2010.

[27] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. ACM Trans. Intell. Syst. Technol., 2(3):27:1–27:27, May 2011.

[28] Fabrizio Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1–47, March 2002.

Biographies

WANG Xiang is currently a Ph.D. candidate at the School of Computer, National University of Defense Technology, China. His main research interests include web mining and information security. Email: xiangwangcn@nudt.edu.cn

ZHANG Zhilin is currently a Ph.D. candidate at the School of Computing Science, Simon Fraser University, Canada. His main research interests include data mining and cloud security. Email: zhilinz@sfu.ca

YU Xiang is currently a post-doctor at the School of Computer, National University of Defense Technology, China. His main research interests include web mining and information security. Email: yuxiang@nudt.edu.cn

JIA Yan professor with the School of Computer, National University of Defense Technology, China. Her main research interests include data mining and information security. Email: jiayan@nudt.edu.cn

ZHOU Bin professor with the School of Computer, National University of Defense Technology, China. His main research interests include text mining and information security. Email: binzhou@nudt.edu.cn

LI Shasha received the Ph.D. degree at the School of Computer, National University of Defense Technology, China. His main research interests include text mining and information security. Email: lishasha198211@163.com