Detecting Spam on Sina Weibo

Yingcai Ma
School of Computer Science and Technology, Harbin University of Science and Technology
Harbin, 150080, China
mayingcai88@gmail.com

Yan Niu, Yan Ren
CNCERT/CC China
Beijing, 100029, China
niuyan@cert.org.cn,
renyan@cert.org.cn

Yibo Xue*
Tsinghua National Laboratory for Information Science and Technology
Beijing, 100084, China
yiboxue@tsinghua.edu.cn

***Abstract*—Online social network becomes greatly prevalent and evolves a communication channel for billions of users. Unfortunately, due to the ease of reaching these users, it has been penetrated by spammers who post inappropriate content. After revealing the transmission mechanism of the spam, an automatic detecting framework is designed to identify spam information. The profiles which have multiple discriminative features are extracted for the Machine Learning techniques. In the experiment phase we collected 562K messages posted by 28,679 users on Sina Weibo, then analyzed the different behaviors between malicious accounts and normal ones. We evaluate our approach on a real large-scale dataset. The results demonstrate the effectiveness of the detecting system.**

Keywords-Spam Detection; Social Network Security; Machine Learning

I. INTRODUCTION

Today, online social network (OSN) is one of the most widely used styles of communication. Billions of users spend considerable time on posting their track via online social networking sites, such as Sina Weibo, Twitter and Facebook. OSN can construct an intrinsic, trustable relationship network based on sharing, propagating and distributing information. For example, users of Sina Weibo can build individual community by posting messages (limited up to 140 characters) and instant sharing, which is called *weibo*.

In OSN, information across the social network spread rapidly and effectively in contrast to traditional e-mails. The goal of OSN is to allow friends communicate and keep in touch by interchanging short messages. However, tremendous interacts also involve the malicious information. These messages often contain offensive contents, vicious links and hacker actions, or induce victims to click the phishing website. In 2012, 144 billion of social interactions were produced per day worldwide, among them 68.8% were spam who disseminated abusive electronic content [1]. It is necessary to combat with abusive information

In this paper, we choose Sina Weibo as the battlefield. Currently, it is the most popular micro-blogging site in China, and has over 500 million registered users as of Dec 2012. About 100 million messages are posted every day [2]. Like all other OSN platforms, Sina Weibo has witnessed a variety of spam attacks. The radius of Sina Weibo is very short, which make it difficult to identify spam by traditional measures. An automatic detection framework is proposed to filter spam accounts. We have performed data collection for half a month between 4/20/2013 and 5/6/2013, excluding 5/1/2013 (we have not traced on this day). After analyzing discriminative natures of spam, Machine Learning techniques are applied for detecting spam accounts. The major contributions are as follows:

- A social graph model is depicted to explore the constitution of OSN. The transmission mechanism is described based on Weibo's policy;
- We developed a data collector to obtain the real-time data by Sina API [3]. The dataset consist of 562K weibos posted by 28,679 users. Spam behavior is captured from the dataset;
- We integrate support vector machine, random forest, naïve bayes algorithms for large scale data. A scalable detecting framework is proposed to identify spam with limited human efforts;
- Many previous researches mainly focus on Twitter, Facebook or other popular microblogs. In contrast to these English microblogs, it is scarce for the study of Chinese microblogs. We devote the spam detection of Sina Weibo and validate our approach to real Chinese data;
- Analyze the dataset and assert the experiments to confirm the effectiveness of the proposed framework.

The remainder of the paper is organized as follows. In Section II we discuss related work. Section III provides the graph model on OSN. Spam behavior is described in Section IV, which formalize features we extracted. Section V presents the experiment and evaluates the prediction results. In the end, we briefly discuss the potential issues and make a conclusion of the paper.

II. RELATED WORK

We investigate the related work on social spam detection. To the best of our knowledge, the existing work mainly focuses on traditional e-mail filtering [4] and web spam detecting [5]. Recent researches about OSN for spam detection are still not mature and not much literature related to the subject exists.

Bosma and Meij [6] use a variation of the HITS link analysis algorithm based on user spam reports. Wang [7] introduces a novel content-based features and graph-based features to facilitate spam detection. Bayesian classifier has the best overall performance in term of F-measure and achieve 89% precision. In [8], the authors develop the collective detection approach by combining both content and behavior to distinguish spam campaigns from legitimate ones.

*Corresponding author. Tel: +86 010 62772393. E-mail: yiboxue@tsinghua.edu.cn

Our work shifts the perspective from English microblogs to Sina Weibo and focuses on detecting spam accounts. Multiple novel features are collected from spam.

III. Social Graph Model

OSN comprises a mass of users which are closely connected with each other. In our study, a representative directed graph G = (V, E) is portrayed in Fig. 1, which composes of nodes V (user accounts) and edges E (connect nodes). Every edge stands for relationship between users. Following someone means subscribing their weibo as a reader. If node (user) C follows node B, C is B's follower and B is C's following. C will receive real-time messages as soon as B post weibo. In addition, C does not have a mandatory requirement to approve or follow back. Therefore, OSN is a directed graph whose edges is delegated to following and follower. It is noteworthy that the graph is bi-directional on A and B, namely, a tight connection has been set up since they are mutual friend. Another peculiarity is that a message can be pushed repeatedly along the track of the social graph. If C post a message, via B and A it can then be forwarded to D.

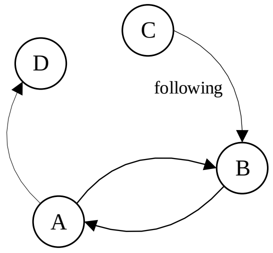


Figure 1. Social Network Graph

IV. Spam Detection Framework

We reveal the behaviors of spam accounts. The purpose is to leverage spam activity for uncovering spam in the wild. But no single nature is effectively capable of distinction between legitimate and spam ones.

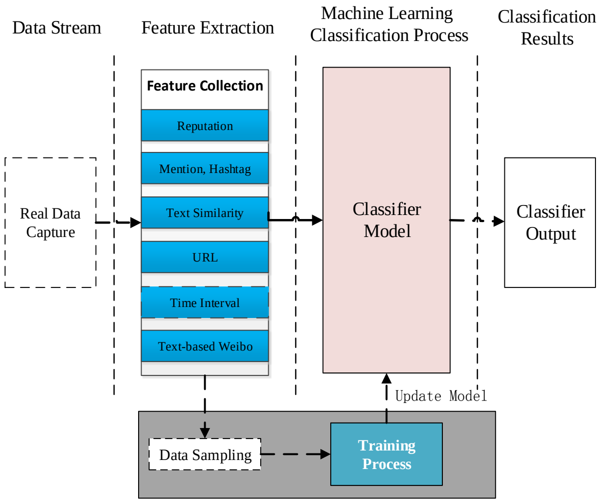


Figure 2. Overview of the Spam Detection Framework

In this section, multiple discriminative features are collected from user accounts, which later will be applied on the machine learning algorithms. We present the social spam detection framework. An overview of the framework is shown in Fig. 2 and we present the two main parts in the following subsections.

A. *Feature Collection and Extraction*

1) Following, Follower and Reputation

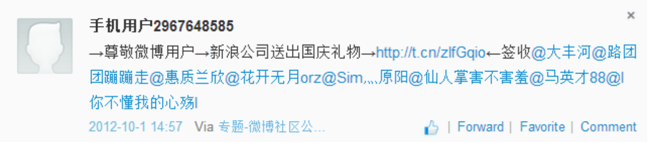
Nearly all OSN sites allow their users to report the suspicious messages or accounts. In order to inject malicious information continuously, spammers make themselves as a true man. Keeping a far-ranging social friendship might convince more people to follow and maintain the contacts. A spammer isn't an actual person so that nobody know his/her existence. Only a part of users agree with their friend requests. On the other hand, the spam account is following others without break. If an account has a small amount of followers compared with a mass of followings, it is suspected to a spammer. A following action indicates out-edge and a follower relation indicates in-edge in the social graph model. The number of followings, the number of followers and the reputation of each account are calculated to describe the characteristics. *N*(follower /following) is the count of follower/following, The reputation ratio is defined as follows:

$$Reputation = \frac{N(Follower)}{N(Follower) + N(Following)} \tag{1}$$

2) Mention and Hashtag

The *mention* service allows placing @username even without the existing social relationship. It is an efficient and rapid way to share information with others. The site will collect such requests and inform the receivers. Normal users rarely follow malicious accounts. So the spammers leverage this function that can allure the victims' visit. The mention is an effective measure which can deliver the messages to targeted users directly. Trick spammers often mention many accounts in their posts (see Fig. 3). For every weibo, the number of mentions is regarded as a feature.

Like '#' symbol, the *hashtag* service allows grouping weibo by trend topic. In case the message has been tagged, it will be displayed on the corresponding topic page, even all the users obtain it immediately. On Sina Weibo, sidebar post the *Hot Topics* every day. The spammer also sneaks into popular topics by signing a hot hashtag, which is unrelated spam. Chu et al. [8] call it hashtag hijacking. Depend on remarkable topics, the tendentious spam maybe draw up to attract normal users to read the messages. For every weibo, the number of hashtags is taken as a feature.



手机用户2967648585

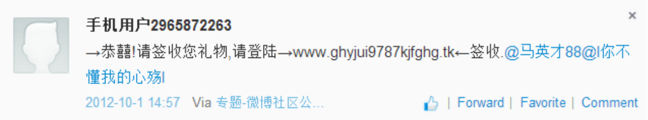
→尊敬微博用户→新浪公司送出国庆礼物→http://t.cn/zlfGqio←签收@大丰河@路团团蹦蹦走@惠质兰欣@花开无月orz@Sim,,,,原阳@仙人掌害不害羞@马英才88@I你不懂我的心殇I

2012-10-1 14:57 Via 专题-微博社区公...

Figure 3. A Sina Weibo Spam Page

3) *URL*

Spammer plots lots of tricky traps for disseminating abusive information and alluring victims to access a harmful or phishing site. Owing to the limit on message length, Sina provides the URL shortening services, which convert all the source links to shorten URL prefixed with “http://t.cn/”. However, this function hides the primitive links and loses sight of the malicious pages behind them. For every weibo, we count the URLs as a feature.



手机用户2965872263

→恭囍!请签收您礼物,请登陆→www.ghyjui9787kjfghg.tk←签收.@马英才88@l你不懂我的心殇l

2012-10-1 14:57 Via 专题-微博社区公...

Figure 4. the Spam Weibo with Disguise URL

Moreover, the tricky scam aim to conceal the feature by transforming URLs. The example in Fig. 4 shows that the spammer don’t comply with the URL shortening services. A remedial measure has been adopted to match this case. If a weibo contains the sequence of characters “http://” or “www.”, it is considered to include a URL by *regular expression pattern.*

4) *Text Similarity*

A spammer may use content templates to deliver numerous weibos, which is a convenient way to post duplicate messages in batch, grouped periodically. But this will cause a high similarity between the messages. A legitimate account often shows mixed characters. The length and the content may not present a strong self-similarity. Thus, the behavior is considered as a feature.

How to calculate the similarity among weibos? Vector Space Model (VSM) is applied to convert text weibo into vector. In each weibo the preprocessing step will remove all mentions, hashtags, URLs to retain pure text. For calculating the similarity, we utilize the Jaccard index, also known as the Jaccard similarity coefficient [9]. It is defined as the size of the intersection divided by the size of the union.

5) *Time Interval*

The spam account isn’t a true person. A large mount of spam accounts are registered by spam campaigns. Those weibos were posted at almost the same time. It looks more likely an automatic tool of sending weibos that was developed by the spammers. A normal user almost do not post a dozen messages in a short period. We only extract the 20 most recent weibos of each account, then calculate the mean and variance of time-interval between two successive weibos.

6) *Text-based Weibo*

In e-mail classification researches, content-based filtering is a reliable method of combating spam, which the reveals some crucial information. Although Sina restricts the maximum length of weibo within 140 characters, the lexical features can still provide useful information.

One simple feature space is the word-based feature space. Sometimes, the spam trick obscures the text by inserting punctuation, and intentional misspelling. All the scenarios give rise to the failure of this method. Here a *inexact string matching* is employed, which is 3-grams using overlapping 3 characters. It were first developed to measure similarity among mutating genes in computational biology [10].

B. *Spam Detection Classifier*

The features are described in the previous will be preprocessed and extracted for machine learning algorithms. Support vector machine (SVM), random forest (RF) and na ïve bayes (NB) is deployed for spam detection.

1) *Support Vector Machine*

SVM has be advocated for classification problem [11]. It has often been shown to give a state-of-the-art performance, which generates a soft margin hyperplane in data space.

2) *Random Forest*

RF is an ensemble classifier that consists of many decision trees. RF tries to improve on bagging by “de-correlating” the trees. Each tree has the same expectation [12]. It predicts the category that is the vote of the classifiers' output by individual trees.

3) *Na ïve Bayes*

A NB classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. NB is light-weighted, fast, robust, and easy to implement. The algorithm is widely applied in classification problems.

V. EXPERIMENTS

A. *Data Collector*

We developed a data collector for testing and evaluating the spam detection framework. 100 seeds are randomly picked up, except for some of them are spam seeds by manual injection. The followings, followers and 20 most recent weibos of each account are captured from 4/20/2013 to 5/6/2013, excluding 5/1/2013(we do not have trace on this day). We gathered more than 562K weibos posted by 28,679 accounts. At last, we manually labeled each account whether it's a spammer or not. In total, 2386 spam accounts and 46K spam messages were checked out.

B. *Data Analysis*

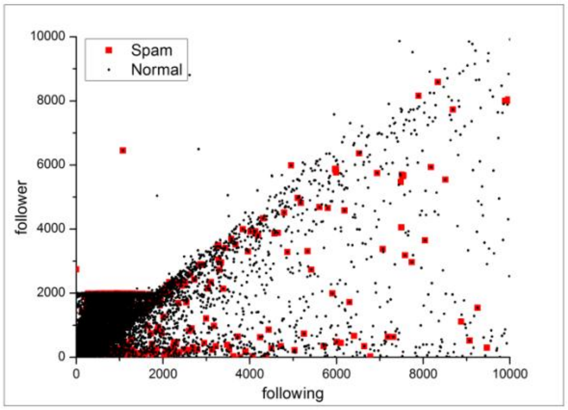


Figure 5. the followings, followers between spam and normal users

Fig. 5 points out the distribution of followings, followers between spam and legitimate users, with follower on the vertical axis and following on the horizontal axis. If an account has a mass of followings compared to the small amount of followers, she/he is suspected to a spammer. But as is shown in the figure that is not always the case. Even though a spam account has no any following or follower, the *mention* and *hashtag* function can be utilized to deliver spam. The spammer can post harmful messages on his own homepage, and mention arbitrary user by the @username format. The notice will be prompted on victims' tab, then induce them to click the link. It is a cheap and efficient way to send spam.

Other features were analyzed and summarized in Table I. All the features are taken on average values. The first four lines were calculated to each weibo, and the time-interval directed at every account base on the minutes scale. Obvious differences appeared in the features, which can help us to segregate spam users from the normal.

TABLE I. FEATURE PROPERTIES OF DATASET

Feature	Spam	Normal
Reputation	0.39	0.12
Mention	9.74	1.28
Hashtag	0.14	0.06
URL	0.89	0.23
Text Similarity	0.79	0.17
Time Interval (Mean)	3.6	214
Time Interval (Variance)	2.8	189

C. Evaluation

The experiments are run with *10-fold cross validation*. The dataset is partitioned into ten non-overlapping parts of equal size. Each part is then tested using the other nine parts of the data as the training data. The final evaluation is averaged over the 10 classification process.

We use the evaluation measures developed for TREC spam track including precision, recall and ROC Area. The area under the ROC curve indicates the probability that a random spam user will receive a lower *spamminess score* than a normal user [4]. For consistency, ROC Area will be reported as the standard performance measure, where 1.0 is optimal.

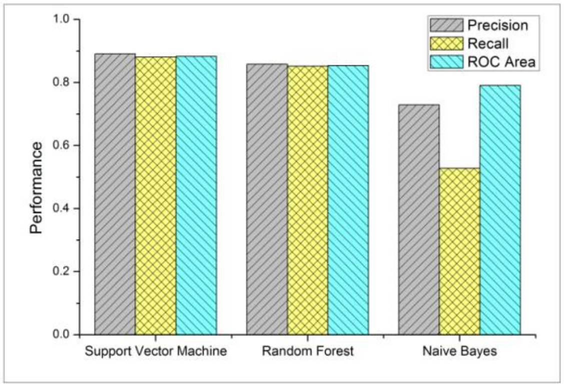


Figure 6. Algorithm Performance Comparison

We performed three popular machine learning algorithms on detection model. As shown in Fig. 6, Note also that they achieve different performance levels. NB classifier is obviously the weakest compared with SVM and RF. SVM gives the strongest performance, and has a slight head than RF. The results indicate the effectiveness of our approach in spam detection.

VI. CONCLUSION

In the paper, we study the transmission mechanism on online social network. The spam activity has been revealed for leveraging spam flooding. Moreover, various distinguishing properties are gathered for detecting malicious information. We developed a data collector and implemented an automatic spam detection framework. Apply the system to real large-scale data. The experiment shows the effectiveness of our approach. The limitations of the data sets do not allow us to draw comprehensive profiles about spammers, such as *user spam reports*. In the future we will investigate a more principled scheme for detecting spam.

ACKNOWLEDGMENT

This work was supported by the National Key Technology R&D Program of China under Grant No.2012BAH46B04. We would like to thank the reviewers for their insightful comments.

REFERENCES

[1] Internet Archive: Pingdom Available: http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

[2] Wikipedia, "Wikipedia: Sina weibo," 2013, [Online; accessed 2-May-2013]. [Online]. Available: http://en.wikipedia.org/wiki/Sina_Weibo

[3] Sina Weibo open platform, development API: http://open.weibo.com/development/pro

[4] Cormack G V, Lynam T, "TREC 2005 spam track overview," The Fourteenth Text REtrieval Conference Proceedings, 2005.

[5] Castillo C, Donato D, Gionis A, et al. "Know your neighbors: Web spam detection using the web topology," Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 423-430.

[6] Bosma M, Meij E, Weerkamp W, "A framework for unsupervised spam detection in social networking sites," Advances in Information Retrieval. Springer Berlin Heidelberg, 2012: 364-375.

[7] W. Alex Hai, "Don't follow me: Spam detection in twitter," Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on. IEEE, 2010.

[8] Chu Z, Widjaja I, Wang H, "Detecting social spam campaigns on twitter," Applied Cryptography and Network Security. Springer Berlin Heidelberg, 2012: 455-472.

[9] Jaccard P, "The distribution of the flora in the alpine zone," New Phytologist, 1912, 11 (2): 37-50.

[10] D. Sculley, G. Wachman, and C. Brodley, "Spam filtering using inexact string matching in explicit feature space with on-line linear classifiers," In The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings, 2006.

[11] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," In the Proceedings of the 10th European Conference on Machine Learning, pages 137–142, 1998.

[12] Breiman L, "Random forests," Machine learning, 2001, 45 (1): 5-32.