The Weibo Spammers' Identification and Detection based on Bayesian-algorithm

Yingying HUANG[1, a], Mengyi ZHANG[2, b], Yuqing YANG[3, c], ShijieGAN[4, d], Yanmei ZHANG[5, e]

[1]Information School of Central University of Finance and Economics, Beijing, 100081, China

[2]Public Finance School of Central University of Finance and Economics, Beijing, 100081, China

[3]Information School of Central University of Finance and Economics, Beijing, 100081, China

[4]Information School of Central University of Finance and Economics, Beijing, 100081, China

[5]Information School of Central University of Finance and Economics, Beijing, 100081, China

[a]email:Doro_hyy@163.com, [b]email:690385667@qq.com , [c]email:819235823@qq.com , [d]email:1308058625@qq.com , [e]email:jizym0309@sina.com

Keywords: Weibo; spammer; identification and detection; Bayesian-algorithm; Genetic-algorithm;

Abstract. The Weibo spammers are employed in latent network public relations or marketing companies, whose job is to achieve publicity or vilify the effect by means of organizing and planning the focus of speculation to a topic or person. In order to reduce the bad influence caused by them, this paper intends to establish a classifier based on the behavior characteristics. By analyzing the previous research, we set the ratio of followers, total number of blog posts, the number of friends, comprehensive quality evaluation and favorates according to latest data. Based on Bayesianalgorithm and Geneticalgorithm, we use R and Matlab to determine the optimal threshold matrix and conditional probability matrix of the changeable Weibo spammers. After testing, it has higher recognition accuracy.

Introduction

With the widespread application of mobile Internet, Weibo has become an important way to gain information and make statements. While it also creates a fast growing number of spammers, who tend to publish tendentious remarks which confuses the public and the authenticity of information.

The Weibo spammers are employed in latent network public relations and marketing companies, who-se job is to achieve publicity or vilify the effect by means of organizing and and casing a focus of speculation to a topic or person.

They try to guide the public opinion on Weibo, so that people's real needs cannot be expressed, which leads to 'one-sided' speech. In addition, spammers jeopardize the network order, citizens' interests, social stability and even national security. Therefore, focusing on the fight against spammers and identifying them on the social network can help control public opinion and reduce false information, hence to refine the effective data as well as purify network environment.

This paper is based on the Bayesian-algorithm, supplemented by the genetic-algorithm. In particular, we crawled large amounts of data from the Weibo platform, using followers, the number of friends, registration time and other aspects to assess and analyze a Weibo user, which leads to higher recognition accuracy.

Related Work

Many recent spammer detection methods have been proposed in literature, behavior and relationship(such as interaction analysis and relationship graph).In earlier studies, the detection of spammers is mainly based on content features, which involves machine learning in natural language processing branch, including text analysis [1], tendency analysis [2] and sentiment analysis [3], etc., using an algorithm such as text analysis, keyword classification, B-Tree indexes[4] to identify the similarity and tendency of comments. Since spammers are better at hiding themselves, methods only based on contents often miss those who spread with normal text, which lead to lower practicability. For behavior detection approaches including Bayesian algorithm [5], decision tree [6], k-means [7] and logistic regression algorithm, they use some features as attributes. However, it's difficult to adapt to the changing and self-disguise of new Weibo spammers. Besides, another drawback of these methods is the lack of a clear threshold to distinguish between spammers and legitimate users. There are also some based on the features of the user's relationship, such as neural network classification, Bias network. Although this kind of method combines with nodes and edges, it needs a lot of complex data to train a network, which is not easy to operate.

Users on some social platforms can be well classified by identifying typical behavior features. In this regard, Parameswaran et al. [8] find that Spammers have unstable way of behavior and come up with the idea to monitor different users for long-term, and then they move the detected Spammers to black list. Gargari et al. [9] use features that most of the Spammers using similar resource and having similar behavior pattern to reach a higher Spammers' detecting precision. Mo[10] believes that legitimate users may gradually form an user-centered social circle in social networks by interaction, while Spammers have abnormal relationships and may form a special network, which contains unbalanced follower-friend ratio. Krestel et al. [11] use the spread of network model to detect Spammers with Forwarding or comments containing links, aiming to set Anomoly Degree on partial seed nodes, using the feature that these nodes will spread out Anomoly Degree to calculate and identify all suspicious nodes. Based on graph theory, Gayo-Avello et al. [12] use abnormal length of time that Spammers spend in following target users or waiting to be followed by them. In addition, he proposes a topic rank which Spammers care passionately about and also uses the users' influence feature to improve Spammers' precision.

Some of the studies only detect and analyze on one feature, which are not perfect for changeable Spammers. As they are getting more complicated, previous algorithm cannot consider well-rounded features which makes it easy to cause Detection vulnerability. Besides, this kind of machine is probably error in collecting few legitimate users with implicit features. Though it may return better Recall and better Precision, it still cannot make a comprehension analysis on Spammers' behavior. Hence these are not suitable in long-term research.

Weibo Spammers' Identification And Detection Modeling Based On Bayesian-Algorithm

There are three main methods in detecting Weibo Spammers: an approach based on content and sentiment, an approach based on probability graph and an approach based on network interaction. Among all, analysis based on content and sentiment requires long-term observation on Spammers to extract typical features. Moreover, this could only simply distinguish users by text information that has been published. As we can see, the rapid growth of Spammers makes their text information more closely to those released by Legitimate Users, which causes a performance decline to the approach. Likewise, though detection based on sophisticated network that formed by interaction extracts features of nodes and edges (represent users and their relations respectively) which can improve accuracy of identification, there is still trouble in crawling data and this is not beneficial for real-time information analysis (which could affect the result of classification as well). To tackle problems above, this paper presents an approach based on Bayesian Algorithm that combines Probability Graph, also using Genetic Algorithm to create the optimal threshold matrix to compensate previous works' lack of incapability to update the adaptive threshold matrix. By this mean, the classifier could have better adaptability to reality to detect Spammers with higher precision.

Two kinds of identifying model will be proposed to fit different types of attributes in this study. Bayesian Algorithm will be based to gain the conditional probability matrix by measuring the occurrence of the same set of attributes in different threshold. Then Genetic Algorithm will help to find the optimal threshold matrix. At the end, confusion matrix will be used to fix the problem that the numbers of two samples (Spammers and Legitimate users) could not match and then return an accurate precision which can reflect the detecting capacity of the classifier appropriately. In this case, the major works are presented as follows:

A. Set Representative Attributes

According to the circumstance that Spammers are reproducing in Weibo, typical features adapting to the reality will be set as attributes. Instead of analysis based on content and sentiment, objective data like Followers' number, Friends' number is easier to update and process. Other four features in personal profiles are given with different weights to create a more representative figure named "Quality Evaluation(QE)".In addition, "sunshine credit" is added as a latest data which is also a new value set in Weibo at the end of 2015.

B. Spammers' Detection Based On Bayesian Algorithm and Genetic Algorithm

Bayesian Algorithm and Genetic Algorithm are involved to build Spammer Judgment and Threshold Optimization models. Then we use the optimal threshold matrix and the conditional probability matrix to detect Spammers by input appropriate attribute parameters.

C. Train Models with Different Sources of Data

Two model will be built to fit different kinds of attribute parameters (here use data from different resources to increase classifier's credibility and comparability). Each set of data crawled and processed by different means and can be divided into two groups (data processed by Manual Annotation and data crawling from known-types collection). Two models are built in accordance with the same idea but different dataset which makes it more accurate when facing different attributes.

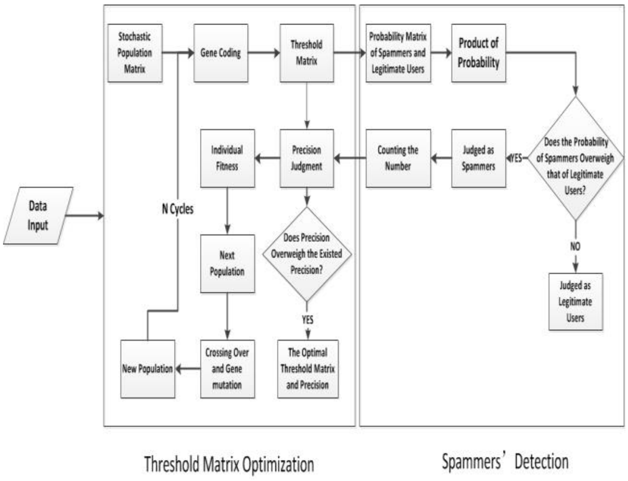
Stochastic Population Matrix
Gene Coding
Threshold Matrix
Probability Matrix of Spammers and Legitimate Users
Product of Probability
Individual Fitness
Precision Judgment
Counting the Number
Judged as Spammers
YES
Does the Probability of Spammers Overweigh that of Legitimate Users?
Data Input
N Cycles
Next Population
Does Precision Overweigh the Existed Precision?
NO
Judged as Legitimate Users
YES
New Population
Crossing Over and Gene mutation
The Optimal Threshold Matrix and Precision
Threshold Matrix Optimization
Spammers' Detection

Fig. 1. Overview of the Classifier's Framework

Spammers' Identification and Detection Algorithm Based on Bayes

A. Basic Attribute Set

By analyzing the previous research and recent changes of spammers, the attributes used in this paper are given as follows:

FF: the ratio of a user's fans to its followers.

FF=Fans/Followers

AW: the average number of Weibo posts per day after registration

AW=Weibo_numbers/Days

IF: the ratio of friends to its followers

IF=Interact/Followers

QE: involves four attributes, which are whether there are comments, brief introductions, authentication and a user's level, and the weights of those are 0.2,0.3,0.3,0.2.

QE=0.2E+0.3R+0.3I+0.2A

During the process, we choose the latest five comments to make sure the data is simpler and the attribute is more distinguished.

C: the number of favorites.

Sunshine credit: Sunshine credit is a new attribute, divided into five levels which are very low, low, average, high, and extremely high. We use 1, 2,3,4,5 to represent those in the experiment.

B. Spammers' Judgment Algorithm

Spammers' Judgment Algorithm is mainly based on machine learning and Bayes theorem.

After setting basic attributes, a threshold matrix of Legitimate users(M) with i rows and j columns(i represents the number of attributes, j represents threshold of the corresponding attribute) must be established. The number of Legitimate Users is n1, which is collected by the manual annotation. Then we need to establish an initial probability matrix of Legitimate Users(T) with i rows and j columns(T_{ij} represents the probability of the i-th attribute falls between M_{ij} and M(i,j+1)). All these probabilities are obtained by the method of counting statistics. Similarly, we use the same method to establish a threshold matrix(N) and an initial probability matrix(S) of spammers(the number of spammers is n2 and S_{ij} represents the probability of the i-th attribute falls between N_{ij} and N(i,j+1)).

Data set X(X = $\{a_1, a_2 \ldots\ldots a_{m-1}, a_m\}$) is not classified, and "a"represents the value of each attribute. Category set B:B = {y_1,y_2}(y_1 represents this is a Legitimate user, while y_2 represents this is a Spammer).The detection result is the larger of the P(y_1|x) and P(y_2|x).

According to Bayes theorem, P(y_i|x) is given by equation 1.

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)} \qquad (1)$$

The attributes of each feature are independent, and the denominator is a constant, besides, the number of spammers and legitimate users in the test data set is clear, which means P(y_i) is also clear. And because of that, when the maximum value of P(x_i|y) is obtained, P(y_i|x) can get its maximum value. So P(x_i|y) P(y_i) is given by equation 2.

$$P(x|y_i)P(y_i) = P(a_j|y_i)P(a_j|y_i)\ldots\ldots P(a_j|y_i)\, P(y_i) = P(y_i)\prod_{j=1}^{m} P(a_j|y_i) \qquad (2)$$

P(a_j|y_i) represents the probability of data a_j's j-th attribute as y_i category.

During the procedure, we add one to the count variable (result) only when it is considered as a spammer. So the total number of spammers is stored in 'result'.

Then, the threshold matrix is optimized by the genetic algorithm. Here we omit the analysis of pseudocode and algorithm complexity.

Experiments and Analysis

A. Experimental Parameters Set

The classifier creates an i rows (i represents for attributes 'number) 4 columns threshold matrix(named var) during optimizing the threshold matrix with Genetic Algorithm. The range of attributes' value is divided into three parts and each columns of the matrix represents partial range. More narrowly, $var_{i,1}$=0， $var_{i,3}$, $var_{i,4}$ stand for the maximum in Spammers, the minimum and the maximum in Legitimate Users respectively. Hence, $var_{i,2}$ is the most valuable threshold optimized by the machine. Other constant value will be processed according to experience, like weights in QE. It is a combination of several features with their weights set by experience (e.g., Higher contributed feature will be delivered with higher weight).

In particular, proportionality constant of Spammers and Legitimate Users used in Spammer Judgment module are calculated from training set. Besides, confusion matrix will be used to tackle the situation that the distribution of Spammers and Legitimate Users differ greatly to adjust the result ran by the classifier.

B. Data Set

As Bayesian Algorithm is based in our study, prior probability of Spammers and Legitimate Users (e.g., a classified dataset) are needed. However, the open-source database is out of date. Some of the data even cannot be used to describe the changeable Spammers any more. Thus a latest dataset is needed in training the detection model. In this circumstance, control experiment is implemented. The process is divided into two groups according to different dataset(i.e., one crawled by R and classified by Manual Annotation , the other consists of Spammers purchased from sales platform and Legitimate Users crawled from friends and relatives), using different attributes to train the classifier.

1) Manually Annotated Data Collection

As a statistical Analysis tool, R is used to crawl about seventy thousand (70,000) data in datapool in different time periods by using Rweibo package. Then we invited a large number of volunteers to tag all the data. After that we integrated all the result and tagged every single data with the type that has the highest recognition. Finally we gained more than a thousand Spammers data and the rest of the Legitimate Users data to train and test the classifier as the experimental group.

2) Confirmed Data Collection

By means of purchasing Spammers from sales platform, we crawled 600 Spammers data. Similarly, we crawled 400 Legitimate Users data from team members' friends and relatives. Both constitute the training and testing data of the control group.

C. Result Analysis

1) Evaluation Criteria

Precision

To improve the accuracy of our classifier, considering this paper using two imbalanced datasets, we build an confusion matrix[13]: TP(True Positive) represents the quantities which are predicted as spammers in spammer samples; TN(True Negative) represents the quantities which are predicted as legitimate users in legitimate user samples; FN(False Negative) represents the quantities which are not predicted as spammers in spammer samples; FP(False Positive) represents the quantities which are not predicted as legitimate users in legitimate user samples.

TABLE I: Definition of confusion matrix

	Actual Spammer Class	Actual Legitimate User Class
Predicted Spammer Class	TP	FP
Predicted Legitimate User Class	FN	TN

In addition, 'acc+' represents the precision to spammer samples, and 'acc-' represents the precision to legitimate user samples. We use 'g'to represents average classification precision of the classifier, which can only get greater value when both 'acc+' and 'acc-' are higher.

$$\mathrm{acc}^{+} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (3)$$

$$\mathrm{acc}^{-} = \frac{\mathrm{TN}}{\mathrm{TN+FN}} \qquad (4)$$

$$\mathrm{g} = \sqrt{\mathrm{acc}^{+} \cdot \mathrm{acc}^{-}} \qquad (5)$$

Recall

SR (Spammer Recall) = $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$

LR (Legitimate Recall) =$\mathrm{TN}/\mathrm{TN}+\mathrm{FP}$

Figure 2 and 3 reveal Precision and Recall with the change of hereditary algebra in experimental group and in control group respectively.

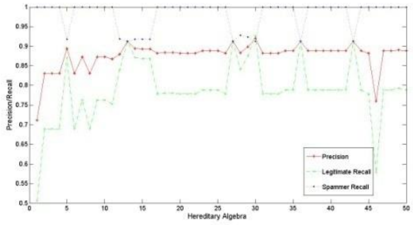

Fig.2. Precision of the experimental group

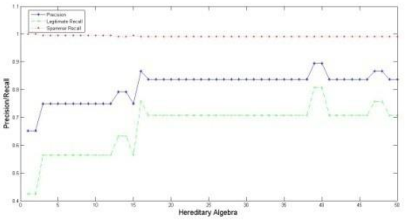

Fig.3. Precision of the control group

F1-Measure

This experiment focuses on F1-Measure of Spammers. Figure 4 and 5 reveal F1-Measure with the change of hereditary algebra in experimental group and in control group respectively.

F1 = 2 *g* SR / (g + SR) (6)

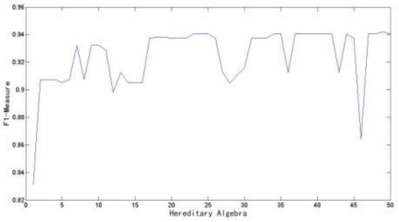

Fig.4. F1-Measure of the experimental group

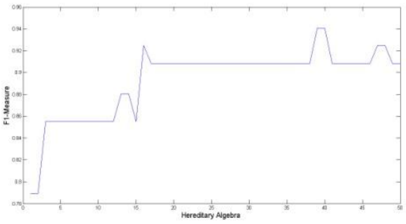

Fig.5. F1-Measure of the control group

2) Comparison and Analysis

Because of the limitation of our data, which is Structured-based and cannot be built into the neural network, the test data will be introduced into the other two commonly used algorithms (logistic regression [14] and decision tree [15]), then calculate the corresponding Precision and Recall. Figure 6 is the performance of different algorithms.

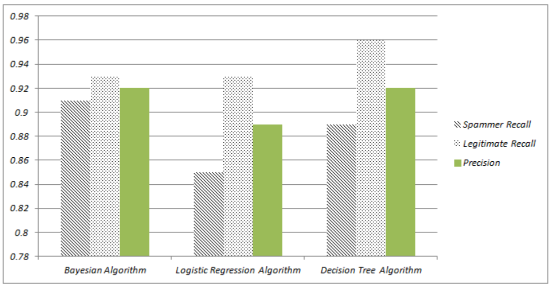

Fig.6.Comparison of accuracy in different algorithms

It is clear that logistic regression and decision tree for Legitimate Recall are lower than that in Bayesian algorithm. Decision tree algorithm has the highest Spammer Recall, and that in logistic regression algorithm and Bayesian algorithm are close. In contrast, logistic regression algorithm has the lowest Precision.

According to these above, there is no doubt that features of Spammers are significant. Thus some legitimate users will be considered as Spammers during the process, which leads to a higher Spammer Recall. Based on linear regression, logistic regression algorithm uses a logistic function. However, in the case of more attributes, a linear relationship is not obvious, so the error is relatively large. Judging through several conditions of data, decision tree algorithm can accurately separate the obvious features of Spammers under the premise of accurate thresholds, but relevant research has not yet given a good definition of threshold selection. Also, it can only recognize the obvious features of legitimate users, hence has a lower Recall.

Overall, Considering on the basis of overall situation, Bayesian algorithm uses a comprehensive measure of the probability of Spammers and Legitimate users to identify a spammer. We also add a threshold matrix with a higher accuracy. This classifier has a higher precision, which guarantees accuracy of the whole algorithm.

Conclusion and Future Work

Nowadays, social networking sites such as Weibo are more frequently used, but at the meantime, the impacts brought by spammers are getting more and more serious. Hence an accurate and efficient spammers' detecting approach has become an urgent need. In this paper, we based on the analysis and summary of related work, using Bayesian Algorithm to build a binary classifier. Starting with the set of typical attributes, the classifier uses Genetic Algorithm to select the optimal threshold matrix and then calculating the conditional probability matrix. Finally, a group of behavior features are set to build a classification model for detecting spammer on Weibo. To build a classifier with higher precision, our study innovatively combines Bayesian Algorithm (which can simultaneously return higher Spammer Recall and Legitimate Recall) and Genetic Algorithm (to create an optimal threshold matrix which can efficiently increase the Precision of the model by improving the conditional probability matrix). From the result, it is easy to see that classifier has a good learning capacity, which shows that it can identify spammers on Weibo effectively. It can also be used to classify Weibo users in practice after training by several latest samples. In future, we aim to provide more extensive attributes and apply the machine to more social platforms.

References

[1] Sriram B, Fuhry D, Demir E, Ferhatosmanoglu H, Demirbas M. Short text classification in Twitter to improve information filtering. In: Crestani F, Marchand-Maillet S, Chen HH, eds. Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2010). New York: ACM Press, 2010.841-842.[doi: 10.1145/1835449.1835643].

[2] Liu B. Sentiment analysis and subjectivity. In: Indurkhya N, Damerau FJ, eds. Handbook of Natural Language Processing. Boca Raton: CRC Press, 2010.627-666.

[3] Zhao YY, Qin B, Liu T. Sentiment analysis.Ruan Jian XueBao/Journal of Software, 2010,21(8):1834-1848 (in Chinese with English abstract).http://www.jos.org.cn/1000-9825/3832.htm [doi: 10.3724/SP.J.1001.2010.03832].

[4] Yang CC, Xu X S, Ye S R, et al. A Method to Find Water Armies in Weibo Based on Text Similarity[J]. Microelectronics &Computer,2014,31(3):82-85.

[5] Cheng X T, Liu C X, Liu S X. Graph-based Features for Identifying Spammers in Microblog Networks[J]. ActaAutomaticaSinica, 2015,41(9):1533-1541.

[6] Liu K, Yuan Y Y, Liu P. A Weibo Bot-Users Indentification Model Based on Random Forest[J]. ActaScientiarumNaturaliumUniversitatisPekinensis,2015,52(2):290-300.

[7] Ni P,Zhang Y Q, Wen GX, et al. Detection of Socialbot Networks Based on Population Characteristics[J].Journal of University of Chinese of Sciences, 2014,31(5):691-700,713

[8] Parameswaran M, Rui H, Sayin S. A game theoretic model and empirical analysis of spammer strategies. In: Proc. of the 7th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conf. (CEAS 2010), Vol.7. 2010.1-7. http://ceas.cc/2010/.

[9] Gargari SM, Oguducu SG. A novel framework for spammer detection in social bookmarking systems. In: Proc. of the IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining (ASONAM 2012). Washington: IEEE Computer Society, 2012.827-834. [doi: 10.1109/ASONAM.2012.150].

[10] Mo Q, Yang K. Overview of Web Spammer Detection[J]. Journal of Software,2014, 25(7):1505-1526[doi:10.13328/j.cnki.jos.004617]

[11] Krestel R, Chen L. Using co-occurrence of tags and resources to identify spammers. In: Saeys Y, Liu H, Inza I, eds. Proc. of the Discovery Challenge Workshop at the European Conf. on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2008). Brookline: Microtome Publishing, 2008. 38-46.

[12] Gayo-Avello D, Brenes DJ. Overcoming spammers in Twitter—A tale of five algorithms. In: Proc. of the Spanish Conf. on Information Retrieval (CERI 2010). 2010. 41-52. http://ir.ii.uam.es/ceri2010/en/.

[13] Pan Z M. Research on Classification for Imbalance Dataset [D]. Xi'an University of Architecture and Technology, 2012:2-49.

[14] Zhang L, Zhu X, Li A P et al. The Spammer Detection based on Logistic Regression [J]. Information Security and Technology, 2015:57-62.

[15] Chen K, Chen L, Zhu P D et al. Interaction based on method for spam detection in online social networks [N]. Journal on Communications, 2015,36(7):120-127.