Detecting Spammers on Twitter Based on Content and Social Interaction

Hua SHEN[1,2]
[1] College of Mathematics and Information Science
Anshan Normal University
[2] School of Software Dalian University of Technology
Anshan Liaoning, China
shenhua_as@126.com

Xinyue LIU
School of Software
Dalian University of Technology
Dalian Liaoning,China
xyliu@dlut.edu.cn

***Abstract*—Twitter has become a target platform on which spammers spread large amounts of harmful information. These malicious spamming activities have seriously threatened normal users' personal privacy and information security. An effective method for detecting spammers is to learn a classifier based on user features and social network information. However, social spammers often change their spamming strategies for evading the detection system. To tackle this challenge, latent user features factorized by text matrix are adopted to capture the consistency of users' behavior. Also, a new social regularization based on users' interaction is introduced to distinguish different types of users. Finally, Supervised Spammer Detection method with Social Interaction is proposed, which jointly learn a classifier by using combine text content, social network information and labeled data. Experimental results on a real-world Twitter dataset confirm the effectiveness of the proposed method.**

Keywords- Social Spammer; Social Regularization; Matrix Factorization

I. INTRODUCTION

Online social networking websites(OSNs), such as Twitter, Facebook and Sina Weibo, have play an important part in people's life. By using these social networking services, it is convenient for people to communicate with their friends easily, publish posts about their life freely, and follow hot topics immediately. One of the most popular OSNs, Twitter, has more than 284 million active users[1,2]. Unfortunately, Twitter has become a new attacked platform for social spammers to achieve their malicious goals such as sending spam [3], spreading malware [4], and performing other illicit activities [5,6]. Therefore, it is important for OSNs to detect social spammers to protect users' privacy, information security and quality of social networking.

Many detection methods have been proposed, which can fall into three categories: features learning method, social-network-based method and optimization method. The first two kinds of methods just consider features learning or social network information, yet the effectiveness of detecting spammer is not ideal. The third method exploits both features as well as social network information to learn an optimized model. However, the design of social regulation could not consider real-world complex phenomena [7]. Consequently, the performances of this kind of methods still need to be improved further.

In this paper, we propose a novel learning model, Supervised Spammer Detection with Social Interaction (SSDSI), which simultaneously integrates social information with content information for detecting spammers on Twitter. Different from the existing methods, the proposed SSDSI takes the frequency of social interaction between users and their neighbors into consideration, which can reflect the real social phenomenon as well as possible. We use matrix factorization technique to induce the latent features of text content. Meanwhile, in order to improve the effectiveness of learning model, we utilize social network information and the label data to guide the latent features learning process. We empirically evaluate the proposed method on a real-world Twitter dataset and describe the advantage of the proposed method.

The rest of this paper is organized as follows. Section 2 reviews the related work on social spammer detection. Section 3 proposes a novel detection model based on content information and social network information. Section 4 presents the empirical results on a real-world dataset. Finally, we conclude this paper and present the future work.

II. RELATED WORK

Social spammer detection has become a hot research topic in academic and industry fields. Many methods have been proposed to identify spammers on Twitter in recent years, including features learning method, social-network-based method, and optimization method.

1) Features Learning Method

Some researchers focus on extracting distinguishing features to train a classifier using supervised machine learning methods. The proposed features can be mainly divided into the following categories: profile-based features, content-based features, graph-based features and neighbor-based features [1,8,9]. However, spammers often change their spamming strategies for evading the detection features. For example, Twitter spammers can purchase a lot of followers from third-party websites or mix many normal tweets to dilute their spam tweets. Therefore, such a phenomenon illustrates that the performance of a detection system depends on classification features only may become less effective over time.

2) Social-network-based Method

Social network information has been used to degrade spammer activities [10,11,12]. These methods utilize random

978-1-4799-1843-0/15 $31.00 © 2015 IEEE
DOI 10.1109/ICNISC.2015.82

walk theory to compute the users' malicious scores and then assist in detecting spammers on Twitter. This kind of methods could not be regarded as a full detection method [10], but rather to be incorporated into the detection system by combining with other detection features.

3) Optimization Method

Some studies have been focused on training classification models with optimization leaning, which consider not only classification features but also social network information [13,14]. The classification features are mostly about users' social activities or content information, that is, these features are hard to evade spammer detection. The social network information is used to collaboratively learn a classification model, rather than to obtain the malicious scores of users. However, a widely used assumption about social network information is that most of spammers' neighbors are normal users and normal users' neighbors are also normal users. In fact, due to spammers' link farms and normal users' courtesy, many of spammers' neighbors are spammers and some of normal users' neighbors are also spammers. Thus, spammer detection is still a challenge for researchers.

III. THE PROPOSED APPROACH-SSDSI

In this section, we first introduce matrix factorization to learn the latent feature matrix on user tweets. and then propose a social regularization with social interaction coefficient to guide the factorization of the latent matrix. Finally, we jointly combine supervised knowledge with matrix factorization and social regularization processes.

A. Modeling Text Content Information

Previous researches have shown that many features, such as profile-based features and graph-based features, are easy to be evaded because of spammers' sophistication and changing. However, text content post by users not only could truly reflect users' intentions, but also it is hard to be evaded. Therefore，text content could be used as features to learn classification model. Considering the representation model of text content is sparse and high dimensional, we first factorize the text content matrix $X \in R^{m \times n}$ into two matrices $U \in R^{m \times k}$ and $V \in R^{n \times k}$, where m is the number of different words, n is the number of users and k is the number of latent features. U and V denote base matrix and coefficient matrix respectively. Coefficient matrix V is regarded as the latent user feature matrix, which is used to learn the classification model. Text matrix factorization is as follows:

$$\min o(U,V) = \frac{1}{2} \| X - UV^T \|_F^2 + \frac{\lambda_f}{2} (\| U \|_F^2 + \| V \|_F^2). \quad (1)$$

The first term is the approximating loss and the latter term is the regularization part to avoid overfitting the factorization. λ_f is a regularization coefficient, which is to tradeoff between the approximating loss and the regularization terms.

B. Modeling Social Network Information with Social Interaction

The existing researches show that social relationships between users are crucial for identifying spammers from social users [13,14,15]. It is obviously that only taking account of text content without considering social network information is not enough for decomposition of the matrix X. Hence, we utilize the social network information to guide the latent features learning process.

Previous works about social network are based on the following assumptions: the normal users perform similarly with their neighbors; on the contrary, the spammers perform differently from their neighbors. In fact, in order to enhance their reputation, spammers construct link farming via creating fake users or purchasing followers, that is, many of spammers' neighbors are spammers. Due to the normal users' courtesy, some of normal users' neighbors are also spammers. Based on our observation, we find that social interaction between users could reflect effectively the similarity of users. In other words, the more frequency of social interaction between users there is, the more likely they are the same type of users, i.e., normal user or spammer.

According to our analysis, social interaction coefficient α_{ij} is defined as

$$\alpha_{ij} = \begin{cases} 0, & \min\{p_{ij}, p_{ji}\} < 3 \\ \dfrac{2 \cdot \min\{p_{ij}, p_{ji}\}}{p_{ij} + p_{ji}}, & \min\{p_{ij}, p_{ji}\} \geq 3 \end{cases} \quad (2)$$

where p_{ij} denotes the number of interaction generated by user i actively to user j, and p_{ji} denotes the number of interaction made by user j actively to user i. From the definition, we can see that $\alpha_{ij} = \alpha_{ji}$ and $\alpha_{ij} \in [0,1]$.

Intuitively, for each user, if its label is same to the label of its neighbor, its latent factors should be similar to its neighbors. Furthermore, the more frequency of social interaction between this user and its neighbor there is, the more similar their latent factors are. Instead, if a user's label is different from its neighbor's, its latent factors should be dissimilar to its neighbors.

Based on the above analysis and the definition of interaction coefficient, social regularization is give as follows:

$$R_s = \sum_{u_i} \sum_{u_j \in N(u_i)} (1 + \alpha_{ij}) y_i y_j \| V_i - V_j \|_2^2 \quad (3)$$

where y_i denotes the label of user i and y_j denotes the label of user j. Vi and Vj denote the latent feature vectors of

user i and user j respectively. If $y = -1$, user is a spammer, and if $y = 1$, the user is a normal user.

C. SSDSI: Supervised Spammer Detection with Social Interaction

Social spammer detection is a classification problem, and so supervised information is a vital factor of detection system performance. Inspired by the Collective Matrix Factorization[16], we then plug supervised information into the above matrix factorization with social information. The popular hinge loss used in Support Vector Machine (SVM) is choosed as the classification model. To use the gradient computation, we adopt the smoothed hinge loss:

$$h(s) = \begin{cases} \frac{1}{2} - s & s \le 0 \\ \frac{1}{2}(1-s)^2 & 0 < s < 1 \\ 0 & s \ge 1 \end{cases} \quad (4)$$

The new optimization objective becomes

$$\min o(U,V,W) = \frac{1}{2} \| X - UV^T \|_F^2 + \frac{\beta}{2} \sum_{i=1}^{l} h(y_i (WV_i^T)) + \frac{\lambda_s}{2} R_s + \frac{\lambda_f}{2} (\| U \|_F^2 + \| V \|_F^2 + \| W \|_F^2) \quad (5)$$

where β is the tradeoff coefficient between the factorization loss and the classification loss, l is the number of labeled data, W is the classification vector for user text latent matrix, and λ_s is the trade off coefficient between the factorization loss and the social regularization.

D. An Optimization Algorithm

We first derive the gradients of each variables in the Eq.(5) as follows.

$$\frac{\partial o}{\partial U} = XV - UV^T V + \lambda_f U$$

$$\frac{\partial o}{\partial V_i} = U^T X_i - U^T U V_i^T + \beta h'(y_i (WV_i^T)) y_i W + \lambda_s \sum_{u_j \in N(u_i)} (1 + \alpha_{ij}) y_j y_j (V_i - V_j) + \lambda_f V_i \quad (6)$$

$$\frac{\partial o}{\partial W} = \beta \sum_{i=1}^{l} h'(y_i (WV_i^T)) y_i V_i + \lambda_f W,$$

where the gradient of the smoothed hinge loss h(s) is

$$h'(s) = \begin{cases} -1 & s \le 0 \\ s - 1 & 0 < s < 1 \\ 0 & s \ge 1 \end{cases} . \quad (7)$$

We use a stochastic gradient descent algorithm(SGD) to optimize the above objective function in Eq.(5). Algorithm 1 is the pseudo-code of the proposed method SSDSI. The input data include text content matrix X, social relation matrix R, interaction matrix P, a part of labeled data Y, the number of latent features K, learning rate and the maximal number of iterations I. The output data are text latent matrix U, user text latent matrix V and classification matrix W.

Algorithm 1
SSDSI: Supervised Spammer Detection with Social

1: Input: Text content matrix X, Social relation matrix R, Interaction matrix P, Labeled data Y, Number of latentfeatures k, Learning rate η and Maximal number of iterations I
2: Initialize U and V
3: For i =1 to I do
4: $U \leftarrow U - \eta \frac{\partial o}{\partial U}$
5: $V \leftarrow V - \eta \frac{\partial o}{\partial V}$
6: $W \leftarrow W - \eta \frac{\partial o}{\partial W}$
7: If convergence break
8: End for
9: Return U,V and W
10: Output: Text latent matrix U, User text latent

IV. Experiments

A. Dataset

We now introduce the real-world Twitter dataset, i.e., UDI Twitter dataset, used in our experiment. This dataset was originally collected in May 2011 on Twitter and introduced in [17]. It contains 140 thousand user profiles, 50 million tweets and 284 million following relationships. We manually scan the tweets content of all users and click the URLs to judge whether they are pornographic information or advertisements. At last, we extracted 1629 spammers and 10450 legitimate users from 12079 users as our dataset, Table 1 shows the statistics of dataset.

TABLE 1. EXPERIMENT DATASET SUMMARY

Dateset	Spammers	Normal users	Tweets	Relationships
Twitter	1629	10450	1087408	740836

B. Evaluation Measures

To evaluate the effectiveness of our proposed model, we use the traditional evaluation measures in social spammers detection, including precision, recall and F1-score. The confusion matrix is shown in table 2, where a is the number of spammer examples that were correctly classified, b is the number of spammer examples that were falsely classified as normal users, c is the number of normal examples that were falsely classified as spammers, and d is the number of normal examples that were correctly classified.

TABLE 2. CONFUSION MATRIX

		Prediction	
		Spam	Normal
True	Spam	a	b
	Normal	c	d

By treating spammers as positive samples in the binary classification, precision is P=a/(a+c), recall is R=a/(a+b), and F1-score is defined as F=2PR/(P+R). Generally, if a classification model achieves higher evaluation metrics, we believe the model is more effective.

C. Performance Evaluation

We compare the proposed method SSDSI with the following baseline methods.

- SVM. In our contrast experiments, Support vector machine (SVM) is employed for spammers detection based on text content information only.
- MF+SVM. We perform the matrix factorization on the user-content matrix, and then use the latent user features to build the classification model.
- SMFSR. It is a matrix factorization method with social regularization based on user activities. The difference with our proposed method (SSDSI) is that this method does not consider the social interaction.
- SSDSI. It is our proposed method with social interaction for spammer detection.

In this paper, we set λ_s=10, λ_f=10, β=100 and K=30. Using these parameters, we adopt 5-fold cross validation to evaluation the effectiveness of the above experiments. In order to explore the effects brought by different sizes of training data, we use two sets of experiments with different numbers of training data, that is, 50% of training data and 100% of training data. For each round of the experiment, 20% of the whole dataset is sampled for testing. "50% of training data" means that we choose 50% of the 80% randomly, thus 40% of the whole dataset is sampled for training. Similarly, "100% of training data" mean that 80% of the whole dataset is used for training. The experimental results of the above methods are presented in Table 3.

From the results in Table 3, we can observe that our proposed method SSDSI consistently outperforms other baseline methods with different sizes of training data. Our method achieves better performance than the state-of-the-art method SMFSR. This indicates that not only our proposed model successfully utilizes both content and social network information for social spammer detection, but also social regularization with social interaction is more beneficial to the learning of classification model than the social regularization proposed by previous method.

TABLE 3. SOCIAL SPAMMER DETECTION RESULTS

Method	50% of the Training Data			100% of the Training Data		
	P	R	F	P	R	F
SVM	0.753	0.821	0.786	0.790	0.848	0.818
MF+SVM	0.781	0.845	0.812	0.812	0.869	0.840
SMFSR	0.824	0.883	0.852	0.856	0.914	0.884
SSDSI	0.837	0.901	0.868	0.868	0.927	0.897

V. CONCLUSION AND FUTURE WORKS

In this paper, we explore the problem of detecting spammers on Twitter. Our proposed method seamlessly integrates feature extraction from text content, social network information and supervised information into a single framework (SSDSI). In particular, our proposed social regularization is different to previous method, which considered social interaction phenomenon between social users on Twitter. The experimental results with a real-world Twitter dataset show that our proposed method is effective and efficient to detect spammers compared with the state-of-the-art methods.

Next, we plan to extend our work in the following aspects. Firstly, we consider other information such as picture message, location and users' sentiment for detecting spammer. Secondly, we wish improve our method and realize a multi-class classification task. Lastly, we will attempt to explore an online detection system, which incrementally update the learning model.

This work was financially supported by National Science Foundation of China (61272374, 61300190), Specialized Research Fund for the Doctoral Program of Higher Education (20120041110046) and Key Project of Chinese Ministry of Education(313011).

REFERENCES

[1] Chao Yang, Robert Chandler Harkreader, Guofei Gu. IEEE: Transactions on Information Forensics and Security, Volume. 8, Issue8,(2013),P. 1280-1293.

[2] Instagram now has more users than Twitter. Information on http://www.trustedreviews.com/news/instagram-now-has-more-users-than-twitter.

[3] Acai Berry spammers hack Twitter accounts to spread adverts. Information on http://nakedsecurity.sophos.com/2010/12/13/acai-berry-spam-gawker-password-hack-twitter/.

[4] New Koobface campaign spreading on Facebook. Information on http://forums.cnet.com/7726-6132_102-5064273.html.

[5] Twitter phishing hack hits BBC, Guardian and cabinet minister. Information on http://www.guardian.co.uk/technology/2010/feb/26/twitter-hack-spread-phishing.

[6] Xinyue Liu, You Wang, Shaoping Zhu, et al. ELSEVIER: Pattern Recognition Letters, Volume. 34, Issue13,(2013),P.1462-1469.

[7] Xinyue Liu, Hua Shen, Fenglong Ma, et al. Topical Influential User Analysis with Relationship Strength Estimation in Twitter. 2014

IEEE International Conference on Data Mining Workshop, 2014:1012-1019.

[8] Benevenuto F, Magno G, Rodrigues T, Almeida V. Detecting spammers on twitter. In Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, Redmond, Washington, 2010: 1-9.

[9] Lee K, Caverlee J, Webb S. Uncovering social spammers: social honeypots+ machine learning. Proceeding of the 33rd international ACM (SIGIR) conference on Research and development in information retrieval, Geneva, Switzerland, 2010:435-442.

[10] Ghosh S, Viswanath B, Kooti F. et al. Understanding and Combating Link Farming in the Twitter Social Network. Proceedings of the 21st international conference on World Wide Web,2012:Pages 61-70.

[11] Chao Yang, Harkreader R., et al. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. Proceedings of the 21st international conference on World Wide Web , 2012: Pages 71-80.

[12] Junxian Huang, Yinglian Xie, Fang Yu, et al. SocialWatch: detection of online service abuse via large-scale social graphs. Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security. ACM, 2013.

[13] Yin Zhu, Xiao Wang, Erheng Zhong, et al. Discovering Spammers in Social Networks. Twenty-sixth AAAI.Conference on Artificial Intelligence. 2012.

[14] Xia Hu, Jiliang Tang, Yangchao Zhang, et al. Social spammer detection in microblogging. Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI), 2013.

[15] Xia Hu, Jiliang Tang, and Huan Liu. Online Social Spammer Detection. Proceedings of the 28th AAAI Conference on Artificial Intelligence, 2014.

[16] Singh, A. P., and Gordon, G. J. 2008. Relational learning via collective matrix factorization. In KDD, 650-658.

[17] Rui Li, Shengjie Wang, Hongbo Deng, et al. Towards social user profiling: unified and discriminative influence model for inferring home locations. In KDD, pages 1023-1031, 2012.