Detecting "Smart" Spammers On Social Network: A Topic Model Approach

Linqing Liu[1], Yao Lu[1], Ye Luo[1],Renxian Zhang[2],Laurent Itti[1,3] and Jianwei Lu[1,4]

1 iLab Tongji, School of Software Enginnering, Tongji University
2 Dept. of Computer Science and Technology, Tongji University
3 Dept. of Computer Science and Neuroscience Program, University of Southern California
4 Institute of Translational Medicine, Tongji University

The Evolution of Spammers

rising of smart accounts in microblog

twitter weibo

Past:

Fake

Legitimate

Now:

simple spammer

smart spammer

Which is fake and which is legitimate?

Introduction

Spammer detection on social network is a challenging problem. The rigid anti-spam rules have resulted in emergence of "smart" spammers. They resemble legitimate users who are difficult to identify. In this paper, we present a novel spammer classification approach based on Latent Dirichlet Allocation (LDA), a topic model. Our approach extracts both the local and the global information (LOSS and GOSS) of topic distribution patterns, which capture the essence of spamming. Tested on one benchmark dataset and one self-collected dataset, our proposed method outperforms other state of-the-art methods in terms of averaged F1- score.

Three Kinds of Users

We observed the difference on topic distribution of different kinds of user.

		Education	Childcare	Shopping Discount	Food	Advertisement
Legitimate Users	Students	0.6	0	0.1	0.3	0
	Parents	0.1	0.6	0.2	0.1	0
Fake Accounts		0.2	0.2	0.2	0.2	0.2
Content Polluters		0	0	0.1	0	0.9

Legitimate users:
- They mainly focus on topics which interest them.

Content Polluters:
- their tweets are all about certain kinds of advertisement and campaign.
-They concentrate on certain topics.

Fake Accounts:
- Their tweets resemble legitimate users but it seems they are simply rando copies of others to avoid being detected by ati-spam rules.
-They focus on wide range of topics.

Latent Dirichlet Allocation(LDA) Topic Model

Each document i is deemed as a bag of words $W = \{w_{i1}, w_{i2}, \cdots, w_{iM}\}$, and M is the number of words. Each word is attributable to one of the document's topics $Z = \{z_{i1}, z_{i2}, \cdots, z_{ik}\}$ and K is the number of topics. ψ_k is a multinomial distribution over words for topic k. θ_i is another multinomial distribution over topics for document i. α and β are hyper parameter that affect scarcity of the document-topic and topic-word distributions.
In this paper, α, β and K are empirically set to 0.3, 0.01 and 15. The entire content of each user is regarded as one document. We adopt Gibbs Sampling to speed up the inference of LDA. Based on LDA, we can get the topic probabilities for all users in the employed dataset as:

$$X = [X_1; X_2; \cdots; X_n] \in \mathbb{R}^{n \times K}$$

where n is the number of users. Each element $X_i = [p(z_1)p(z_2)\cdots p(z_K)] \in \mathbb{R}^{1\times K}$ is a topic probability vector for the ith document. X_i is the raw topic probability vector and our features are developed on top of it.

Feature Extraction

Using the LDA model, each person in the dataset is with a topic probability vector X_i.
Assume $x_{ik} \in X_i$ denotes the likelihood that the ith tweet account favors k^{th}topic in the dataset. Our topic based features can be calculated as below.

Global Outlier Standard Score measures the degree that a user's tweet content is related to a certaintopic compared to the other users. Specifically, the "GOSS" score of user i on topic k can be calculated as below:

$$\mu(x_k) = \frac{\sum_{i=1}^{n} x_{ik}}{n}$$

$$GOSS(x_{ik}) = \frac{x_{ik} - \mu(x_k)}{\sqrt{\sum_i (x_{ik} - \mu(x_k))^2}}$$

The value of $GOSS(x_{ik})$ indicates the interesting degree of this person to the kth topic.

Local Outlier Standard Score measures the degree of interest someone shows to a certain topic by considering his own homepage content only. For instance, the "LOSS" score of account i on topic k can be calculated as

$$\mu(x_i) = \frac{\sum_{k=1}^{K} x_{ik}}{K}$$

$$LOSS(x_{ik}) = \frac{x_{ik} - \mu(x_i)}{\sqrt{\sum_k (x_{ik} - \mu(x_i))^2}}$$

$\mu(x_i)$ represents the averaged interesting degree for all topics with regarding to i^{th} user and his tweet content.

Dataset

We use one public dataset Social Honeypot dataset and one self-collected dataset Weibo dataset to validate the effectiveness of our proposed features.

Social Honeypot Dataset:
Lee et al. (2010) created and deployed 60 seed social accounts on Twitter to attract spammers by reporting back what accounts interact with them. They collected 19,276 legitimate users and 22,223 spammers in their datasets along with their tweet content in 7 months. This is our first test dataset.

Our Weibo Dataset:
Sina Weibo is one of the most famous social platforms in China. It has implemented many features from Twitter. The 2197 legitimate user accounts in this dataset are provided by the Tianchi Competition1 held by Sina Weibo. The spammers are all purchased commercially from multiple vendors on the Internet. We checked them manually and collected 802 suitable "smart" spammers accounts.

Experiment

Feature	Method	Weibo Dataset			Honeypot Dataset		
		Precision	Recall	F1-score	Precision	Recall	F1-score
GOSS	SVM	0.974	0.956	0.965	0.884	0.986	0.932
	Adaboost	0.936	0.929	0.932	0.874	**0.990**	0.928
	RandomForest	0.982	0.956	0.969	0.880	0.969	0.922
LOSS	SVM	0.982	0.958	0.97	0.887	0.983	0.932
	Adaboost	0.941	0.929	0.935	0.878	0.976	0.924
	RandomForest	0.986	0.956	0.971	0.882	0.965	0.922
GOSS+LOSS	SVM	0.986	0.958	0.972	0.890	0.988	**0.934**
	Adaboost	0.938	0.931	0.934	0.881	0.976	0.926
	RandomForest	**0.988**	**0.958**	**0.978**	**0.895**	0.951	0.922

Table 1: Performance comparisons for our features with three baseline classifiers

Feature	Description
UFN	standard deviation of following
	standard deviation of followers
	the number of following
	following and followers ratio
UC	|links| per tweet
	|@username| in tweets / |tweets|
	|unique @username| in tweets / |tweets|
	|unique links| per tweet
UH	the change rate of number of following

Table 3: Honeypot Feature Groups

Features	SVM			Adaboost		
	Precision	Recall	F1-score	Precision	Recall	F1-score
UFN	0.846	0.919	0.881	0.902	0.934	0.918
UC	0.855	0.904	0.879	0.854	0.901	0.877
UH	0.906	0.8	0.85	0.869	0.901	0.885
UFN+UC+UH	0.895	0.893	0.894	0.925	0.920	0.923
LOSS+GOSS	0.890	0.988	0.934	0.881	**0.976**	0.926
UFN+UC+UF+LOSS+GOSS	0.925	0.920	0.923	**0.952**	0.946	**0.949**

Table 2: Comparisons of our features and Lee et al.'s features

Reference

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. the Journal of machine Learning research , 3:993–1022.
[2] Kyumin Lee, Brian David Eoff, and James Caverlee. 2011. Seven months with the devils: A long-term study of content polluters on twitter. In ICWSM . Citeseer.
[3] Xia Hu, Jiliang Tang, Yanchao Zhang, and Huan Liu. 2013. Social spammer detection in microblogging. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence , pages 2633–2639. AAAI Press.
[4] Yu Liu, Bin Wu, Bai Wang, and Guanchen Li. 2014. Sdhm: A hybrid model for spammer detection in weibo. In Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on , pages 942–947. IEEE.