Transactions in GIS

Research Article Transactions in GIS, 2014, 18(3): 370–384

Detecting Non-personal and Spam Users on Geo-tagged Twitter Network

Diansheng Guo and Chao Chen

Department of Geography, University of South Carolina

Abstract

With the rapid growth and popularity of mobile devices and location-aware technologies, online social networks such as Twitter have become an important data source for scientists to conduct geo-social network research. Non-personal accounts, spam users and junk tweets, however, pose severe problems to the extraction of meaningful information and the validation of any research findings on tweets or twitter users. Therefore, the detection of such users is a critical and fundamental step for twitter-related geographic research. In this study, we develop a methodological framework to: (1) extract user characteristics based on geographic, graph-based and content-based features of tweets; (2) construct a training dataset by manually inspecting and labeling a large sample of twitter users; and (3) derive reliable rules and knowledge for detecting non-personal users with supervised classification methods. The extracted geographic characteristics of a user include maximum speed, mean speed, the number of different counties that the user has been to, and others. Content-based characteristics for a user include the number of tweets per month, the percentage of tweets with URLs or Hashtags, and the percentage of tweets with emotions, detected with sentiment analysis. The extracted rules are theoretically interesting and practically useful. Specifically, the results show that geographic features, such as the average speed and frequency of county changes, can serve as important indicators of non-personal users. For non-spatial characteristics, the percentage of tweets with a high human factor index, the percentage of tweets with URLs, and the percentage of tweets with mentioned/replied users are the top three features in detecting non-personal users.

1 Introduction

With the increasing popularity of location-aware devices and platforms, there are more and more user-generated and geo-tagged data sources (Goodchild 2007), which can be classified into two types. For the first type, citizens explicitly provide geographic data, such as roads or geographic boundaries, with examples including OpenstreetMap, Google Mymap, and Wikimapia. For the second type, citizens are willing to turn on their location information when involved in various social activities on different platforms, such as Twitter, Foursquare, Instagram and Flickr. The latter type of data is mostly from location-based social networks, combining social networking services with location technologies such as GPS-enabled mobile devices (Bertrand De et al. 2009).

Twitter is a microblog service that allows users to send messages of at most 140 characters to their followers. As of October in 2013, there are about 231.7 million active users worldwide and 300 billion tweets since Twitter was established. Among all tweets, there are about

Address for correspondence: Diansheng Guo, Department of Geography, University of South Carolina, 709 Bull Street, Columbia, SC 29208, USA. GUOD@mailbox.sc.edu

Acknowledgements: This work was supported in part by the National Science Foundation under Grant No. 0748813. Part of the research is also supported by the Institute of Museum and Library Services (IMLS) through the National Leadership Grant LG-00-14-0030-14.

 doi: 10.1111/tgis.12101

2–3% that have geographic locations, i.e. geotagged. In this research, we focus on the tweets with point locations (i.e. latitude and longitude) and exclude those that only have place names (city, county, or state). For example, for 10-months in the US, there are 5,284,910 active twitter users with at least one geo-tagged tweet, among which 2,935,975 users have at least five geo-tagged tweets with latitudes and longitudes. The analysis of geotagged tweets has proved useful in various areas, such as crisis detection and management (Bertrand De et al. 2009; MacEachren et al. 2011) and location-based marketing and recommendation (Sitaram and Bernardo 2010).

However, not all twitter accounts represent individual people. Driven by various purposes, there are more and more emerging accounts that are organizations, agencies, services, bots, and spammers. The tweets from these non-personal accounts can distort or mislead twitter analysis results in various applications (Benevenuto et al. 2010), especially those on studies of *human* behavior, choice, and activities, which assume that each involved twitter user is an individual person and that his/her tweets represent personal interests and activities. Therefore, it becomes critical to detect non-personal accounts (e.g. organizations, services, bots, and spam users) and exclude them (or treat them separately) in spatial analysis of tweets or twitter users.

In this article, we propose an approach to the detection of non-personal users with geo-tagged tweets on the Twitter network by: (1) extracting content-based features, graph-based features, and geographic features; (2) manually constructing training data sets for supervised classification and knowledge discovery; and (3) evaluating the extracted rules with additional reclassification of new twitter data to understand the characteristics of *non-personal* twitter users. In this research, we broadly define *non-personal* twitter users as those that do *not* post contents on personal daily lives. In general, there are four different kinds of non-personal twitter accounts, including bots (machine accounts that automatically send tweets), organization accounts (such as group or organizational accounts), individual accounts that only post career information, or spam users that post commercial or malicious content. Some of these non-personal accounts and their tweets can be useful services otherwise, such as weather reports and crime alerts. However, they are appropriate for the analysis of spatial and social behaviors of individuals. Therefore, as defined above, we classify organizational accounts and twitter messages as 'non-personal' and separate them from accounts that represent a real person. To our best knowledge, this is the first study that evaluates the characteristics of non-personal twitter accounts within geo-tagged tweets.

2 Literature Review

Spam detection is a classic topic in the study of email systems and web contents (Mishne 2005; Ntoulas et al. 2006). Spam email detection has been studied for a long time (Sahami et al. 1998), with many spam filtering techniques developed and implemented. Compared with email spam, web spam is more challenging to tackle, as it changes rapidly on a large scale. Techniques for detecting web spam can be based on directed graph models (Zhou et al. 2007), link-based and content-based features (Castillo et al. 2007), and semi-supervised algorithms (Geng et al. 2009; Wang 2010a). Meanwhile, spam detection also exists in other fields, such as call spam detection (Wu et al. 2009), video spammer detection, and blog spammer detection (Mishne 2005).

In the past several years, online social networks have become an ideal target for disseminating information, including personal activities, organizational messages, advertisements

or even malicious content (Chu et al. 2012). For example, according to a study (Grier et al. 2010), the probability of a user clicking a Twitter spam is around 0.13%, which is 20 times higher than the probability of clicking an email spam. The reasons for such a high click-through rate are related to several features of the Twitter network and tweets that spam users can take advantage of. First, a tweet message is very short and users often insert URLs to point to target contents such as photos and web pages. The second feature is related to the hashtag, which can be inserted into a tweet to help users identify specific topics. The third feature is the trending topic, which is the most popular and recent of hashtags or keywords in the Twitter network. Spammers can abuse the above features by inserting URLs, identical hashtags or keywords into a large number of tweets. The fourth feature is based on the reply or mention functionality, with which a user can reply or mention other users by adding "@username" in their tweets. The fifth feature is the retweet functionality, which can be used to appeal to potential followers by retweeting other users' tweets. Relationship links on Twitter are directional through the follower network, in which Twitter users can follow other Twitter users (Benevenuto et al. 2010). By analyzing the contents of tweets, one can also extract a friend network to reveal the actual interactions on the Twitter network (Huberman et al. 2009; Yang and Counts 2011). For example, a tweeter user can be deemed as a friend to another user if one is mentioned by another user in tweets using "@username".

Twitter spam detection approaches can be classified into three different types on the basis of the analysis unit, including tweet-level analysis, user-level analysis and campaign-level analysis, At the tweet level, Martinez-Romo and Araujo (2013) applied probabilistic language models to detect the topic divergence of each tweet by analyzing a large number of tweets related to some trending topics. After detecting the topic of a tweet, they calculate the divergence of the topic from the relevant trending topic, using language models, and thus identify spam tweets. In contrast, user-level spam detection has received the most attention with highest classification accuracy. Twitter spam can be detected by studying the tweeting behavior, account age, and network structure of non-spammers and spammers (Yardi et al. 2010). Wang (2010b) provided a prototype to classify suspicious users based on Twitter's spam policy, a number of content-based features, and graph-based features. In a similar work, McCord et al. (2011) applied four traditional classifiers to identify spammers with user and content-based features and the Random Forest classifier. A spam campaign is defined as a collection of Twitter accounts controlled and manipulated by a spammer. Chu et al. (2012) cluster Twitter users into different campaigns based upon the URLs retrieved from the tweets, and then extract features that can be incorporated to classify twitter users.

Unlike previous research on spam detection on the Twitter network, our study focuses on Twitter users who have geo-tagged tweets (e.g. with latitudes and longitudes). Twitter users can send geo-tagged tweets by turning on the location information at various platforms such as the Twitter website, Twitter apps on mobile devices, or other mobile applications (e.g. Instagram and Foursquare). As explained earlier, in this study we are interested in detecting non-personal twitter users who do not post contents related to individuals' daily lives. There can be three different kinds of such accounts: (1) bots, organizations, and user groups; (2) service-oriented accounts that send job-related information; and (3) spammers that distribute commercial or even malicious content. The majority of such accounts are bots with various purposes, including weather alerts, traffic alerts, crime alerts, jobs advertisements, and sales promotions. A list of samples of geo-tagged non-personal accounts is provided in Table 1.

Table 1 A list of sampled geo-tagged non-personal accounts and their tweets

Type	Tweet Sample
Weather alert	Severe Weather Statement issued November 01 at 5:11AM EDT until November 01 at 6:00AM EDT by NWS Blacksburg http://twitzip.com/alerts/294504
Police alert	OrlPolice 32808@orlpol3280821m #ThreatsAssaultsArmed at 2757 Bent Willow Circle. #orlpol
News alert	Man questioned after home explosion in Oak Lawn: Oak Lawn police question a man following a house fire and . . . http://dlvr.it/4hBxbs
Traffic alert	In Guilford closed due to road construction on US-1 NB near Church St. Stopped traffic from I 95
Sales promotion	Single Family, $169000 4 beds 2.1 Baths, 77379 http://www.har.com/87158245RE/MAXIntegrity#Spring
Job advertisement	#CaliforniaJobs, CA #Insurance #Job: Auto Damage Adjuster Trainee at GEICO http://bit.ly/19VVclf#geicojobs #Jobs #TweetMyJobs
Sales promotion accounts	@arrowmediadesig Hi can you promote my company http://imperialimportsinc.com thanks.
Accounts with malicious content	<URL links to inappropriate porn pictures>
Non-individual accounts	AVAILABLE PETS: Bull Winkle http://bit.ly/1bRZ7vQ
Other geotagged spammers	hing. Good thing the tacs don't care about culture or history any more than you do, Captain, or they'd never have put hi

3 Detection of Non-personal Twitter Users with Geo-tagged Tweets

3.1 Overview

In this research, we collected all the geo-tagged tweets for the entire US through the Twitter streaming API. Particularly, we focus on the tweets with point locations (i.e. latitude and longitude) and exclude those that only have place names (city, county, or state). For each tweet we have the account name, latitude/longitude, post time, and the message. We construct training data sets using tweets from a 10-month period (December 2, 2012 – October 6, 2013) for the 48 US contiguous states and Washington DC. The data include 637,330,759 tweets and 5,284,910 unique users, among which 2,935,975 users have at least five geo-tagged tweets with point locations (i.e. latitude and longitude). We extract a collection of content-based, graph-based and geographic features for each user based on all his/her tweets. To ensure the validity of each feature, we only include users that have at least five geo-tagged tweets during that time period. A training data set is constructed through an iterative sampling process with manual inspection and labeling of user types (personal or non-personal accounts). A decision tree is then learned with the training data to extract classification rules for non-personal account detection.

To evaluate the discovered rules, we extract the same set of features for users with tweets from a more recent four-month period (October 7, 2013 – February 8, 2014) for the 48 US contiguous states and Washington DC. This data set includes 314,004,436 tweets and

Figure 1 Screenshot of a legitimate geo-tagged account who occasionally posts advertisements in her non-spatial tweets but has valid geo-tagged tweets through Foursquare check-ins

4,109,330 unique users, among which 2,375,717 unique users have at least five tweets. The discovered rules are applied to classify the 2,375,717 unique users into two categories: non-personal and personal accounts, which we may also call "spam" and legitimate users, respectively. Then we randomly sample a set of predicted non-personal and personal accounts, separately, and manually check the tweets of each sampled user and determine the category. With this re-classification process, we can more reliably assess the robustness of the discovered rules, with data from the different time period that is separate from the training data set.

Before presenting our methodology and results, we would like to emphasize the following two points to ensure correct understanding of our approach. First, in our study we use non-personal accounts and "spammer" interchangeably, referring to non-personal accounts that may be bots, organization accounts, individual accounts that send only work-related information, or spammers that send commercial or malicious content. Figure 1 shows an account that occasionally posts *non-geotagged* tweets to advertise her products; but most of her geotagged tweets are valid tweets sent through Foursquare Check-ins and showing her spatial mobility, thus this account is considered a legitimate (i.e. personal) user in our study. Second, certain types of non-personal users can only be discovered in the geotagged Twitter network by using geographic features such as speed. As shown in Figure 2, the user can be easily identified as a spammer since there is no reasonable transportation to help the user move 7,824 km in one minute.

3.2 *Feature Extraction*

Conventional spam detection schemes on the Twitter network usually use content-based and graph-based features. In this study, we not only extract and use traditional features such as percentage of URLs and percentage of Hashtags, but also derive and use a collection of new features, including geographic features (such as speed and county changes), frequency

Figure 2 A spammer that has invalid locational information (with unreasonable speed)

patterns (e.g. the variance of frequency being mentioned), sentiment-based features (e.g. percentage of tweets with emotions), and an indicator of friendship through replies and mentions.

3.2.1 Content-based features

Our initial detection schema includes 10 content-based features. With the initial classification result, we observed that all users but one who posted through five specific apps were personal users. These five location-based apps are Instagram, Foursquare, Path, Endomondo, and Flickr, all of which provide location-based services such as restaurants check-ins. The tweets posted through these apps contain corresponding external links, which non-spatial approaches tend to treat as spam. Therefore, to improve the detection precision for our definition of non-personal accounts, we take these apps into consideration to refine related features and also propose a new feature called *appTwtsPerc*, measuring the percentage of such app-based tweets of a user's geo-tagged tweets.

Percentage of URLs (***urlPerc***): URLs are widely used in disseminating disruptive contents in the Twitter network. According to Twitter's policy, any user who mainly posts tweets with URLs is considered a spammer. The percentage of URLs is the ratio of the number of tweets containing URL links to the total retrieved geo-tagged tweets for each user. In calculating this feature, we exclude URLs posted through the aforementioned five location-based apps. Otherwise, this feature will be less useful because a certain number of individual users post geo-tagged tweets primarily through these apps, which tend to be identified as spam users due to the extreme high percentage of URLs.

Percentage of Hashtags (***hashtagsPerc***): Hashtag is used to help users to identify particular topics, defined with the format "#symbol" in a tweet. Very popular hashtags become trending topics. This feature is defined as the percentage of tweets that contain hashtags in all retrieved geo-tagged tweets for a particular user.

Percentage of Retweets (***retweetPerc***): Retweets can be abused by spammers by automatically posting retweets. This feature represents the number of tweets containing "RT @username" over all retrieved tweets for each user.

User Sentiment Index (***emotweetPerc***): A sentiment index is proposed to determine whether a user is a bot or a person. Sentiment analysis is used to assess the emotion of each

tweet and assign it an emotion value: positive (1), neutral (0), or negative (–1). The Sentiment140 API, a computing service that can detect sentiment from tweets, is used to conduct the sentiment analysis, which relies on maximum entropy (Go et al. 2009; Nigam et al. 1999). We define the sentiment index for user U_i as follows:

$$S_i = \sum |P_{ik}| / TotalT_i \tag{1}$$

where P_{ik} is the emotion value for the k_{th} tweet of user U_i, and $TotalT_i$ is the total number of retrieved geo-tagged tweets for U_i.

Human Factor Index (***humanFactorPerc***): If a tweet is sent via the aforementioned apps (i.e. Instagram, Foursquare, Path, Endomondo, and Flickr) or has a non-zero emotion score (*emotweetPerc*), we consider that this tweet has the potential to be from a real person and give it a positive human factor score. We calculate a Human Factor Index for each user as the percentage of his/her tweets with a positive human factor score.

In addition to the above mentioned features, we also extract six additional content-based features for each user:

- *maxNumTwtsPerM* – the maximum number of geotagged tweets per month, which is obtained from a 30-day time window for the study period and the maximum value for all windows.
- *minNumTwtsPerM* – the minimum number of geotagged tweets per month in the study period.
- *diffTwtsPerM* – the difference between *maxNumTwtsPerM* and *minNumTwtsPerM*.
- *meanTimeInterval* – the mean time intervals between two consecutive geotagged tweets.
- timeIntervalStd – the standard deviation of time intervals between two consecutive geotagged tweets.
- *percAppTwts* – the percentage of tweets posted via the five above mentioned apps.

3.2.2 Graph-based features

Graph-based features are based on the interaction among users through tweet replies and mentions, which can be used by both legitimate (personal) and non-personal users. We hope to extract features that can find distinctively different patterns for the two types of users.

Percentage of Replies/Mentions (***repliedPerc***): A user can reply to or mention other users by inserting "@username" in tweets. This feature is calculated as the percentage of total tweets for a user that have mentioned or replied someone. Note: the mentioned user in the tweets may or may not have geo-tagged tweets.

Mention/Reply User Count (***uniMentnUsersPerM***): The number of unique users that a user U_i mentioned, normalized with the number of months. Since different users can be mentioned multiple times by U_i, the variance of the frequencies for all mentioned users is calculated, called ***normVar***. We also calculate the number of times that U_i is mentioned by other users per month (***beMntCntsPerM***).

Friend-Network Relationship (***friendIndicator***): A measure defined as the number of unique mentioned/replied users over the number of times of being mentioned by other users (see Equation 2, where 0.01 is added to the denominator to avoid dividing a zero value).

$$friendIndicator_i = \frac{beMntCntsPerM_i}{uniMentnUsers_i + beMntCntsPerM_i + 0.01} \tag{2}$$

3.2.3 Geographic features

We extract seven geographic features for each user. The tweeting speed in our article is defined as the distance between two consecutive tweets of a user divided by the time interval between the two tweets. If the time interval is longer than one day, then we use one day instead. The distance interval is the geographical distance between two consecutive tweets.

- *hasLocsCntsPerM* – the average monthly number of mentioned/replied users who have geotagged tweets. To get this value, we firstly retrieve all the mentioned/replied users for each account based upon his or her geotagged tweets. Then we iterate through all these users and consider one as a user with locations if this user ever sent geotagged tweets. Lastly, we divide the number of users with locations by the number of months to get this variable.
- *uniqueCntyCntsPerM* – the average monthly number of counties in which the user has been.
- *cntyChangesPerM* – the number of times per month that a user moves across county boundaries between consecutive tweets.
- *maxSpeed* – the maximum tweeting speed for a user. To get this value, we find the maximum speed among all of the tweeting speeds between each pair of consecutive tweets that has a distance interval more than 1.6 km (1 mi). The distance threshold is set to avoid speed errors caused by location inaccuracy within a short distance. Most cellular service providers have adopted Assisted-GPS technology that combines GPS (with an average accuracy of 10 m), Wi-Fi (70–80 m), and cellular position (100–300 m) to provide location information (Zandbergen 2009). The distance between two tweet locations can have an error of as much as 500 m or even more when GPS or Wi-Fi signals are not available. Therefore we only calculate speed for distances longer than 1,600 m (i.e. about one mile).
- *meanSpeed* – the average tweeting speed.
- *maxSpeedDist* – the distance interval when the maximum tweeting speed occurs.
- *speedLimitCntsPerM* – average monthly number of times that the user has a traveling speed exceeding a speed threshold (145 km/h), which is about the maximum speed one may drive on most highways. In the process of manual classification, we notice that tweets sent through Instagram and Flickr may be associated with 'wrong' timestamps because users may keep their photos and then update them to Twitter through Instagram or Flickr later. In this case, our computed speed may be higher than the speed threshold and thus may affect our classification. To address this, we ignore the speed for tweets from Instagram or Flickr.

3.3 Manual Creation of Training Data

We construct a training dataset by drawing a sample of users, manually inspecting their tweets and assigning each user a label: personal (legitimate user) or non-personal ("spam"). Due to the large number of Twitter users and the time-consuming process of manual labeling, a random sampling does not work well, as it would require a very large sample size to obtain a sufficient number of non-personal accounts. To disproportionally draw non-personal samples, we take an iterative process. First, we detect suspicious non-personal users with an initial and simple detection scheme based upon the derived speed and the time interval between consecutive tweets. Specifically, we set a time interval limit of two seconds and

a speed limit threshold of 1000 km/hr (which is faster than airplanes) between two consecutive tweets by a user. Through *automatic* processing, a spam label is assigned to a user if they have more than three violations of the two limits, i.e. with a faster speed or a shorter time interval.

After this initial labeling, we randomly select 500 personal users and 500 suspicious non-personal users. Then we *manually* classify these 1,000 users by manually reading each user's top 30 geo-tagged tweets and their top 30 recent non-spatial tweets on the Timeline. Out of the 1,000 sampled users, 204 are actual non-personal users based on our judgment. To increase the diversity and representation of non-personal users, we perform a supervised classification with the above training data set, using the decision tree method and all the extracted user features introduced in Section 3.2. A set of classification rules is obtained, which we apply to label each of the 2,935,975 users in the entire data set. From this labeled data set, we randomly select 100 new "non-personal users" and 100 new personal users. We manually reclassify them by reading their tweets and added 47 non-personal users to the training set. Eventually, the training data has 925 personal users and 252 non-personal users and thus 1,177 users in total. This process can be repeated to add more training samples, in order to obtain sufficient samples of non-personal twitter users.

3.4 Supervised Classification

Four supervised classifiers were applied for the classification task, including Decision Tree (Quinlan 1996), Naive Bayes (Schneider 2003), Support Vector Machine (SVM) (Joachims 1998), and Random Forest (Breiman 2001). In terms of the F-Measure and root mean squared error, Random Forest is the best classifier. For each model, a 10-fold cross-validation was carried out to assess the classification accuracy. For each run, the training data set is randomly split into 10 subsets, of which nine subsets are used together to train the classifier and the remaining subset is used to test the classifier. Each subset is used exactly once for validation. Random Forest is an ensemble method, using many decision tree models, which selects a subset of the dataset with replacement to train each tree, estimates error and variable importance using the remaining dataset, and assigns each user a class based on votes from all trees (Breiman 2001).

A number of evaluation metrics are applied to compare different models, as shown in Table 2. Random Forest outperforms other models with the highest F-measure, highest Precision, highest Recall and lowest root mean square error. Decision Tree has the second best performance in terms of F-measure and accuracy. The disadvantage of Random Forest, however, is that it is more like a black box and does not provide explicit rules. In this regard Decision

Table 2 Evaluation metrics of supervised classifiers

Evaluation Metrics	Recall	Precision	F-measure	Accuracy	Root mean square error
Random Forest	0.959	0.959	0.958	0.959	0.1926
Decision Tree	0.955	0.954	0.954	0.955	0.2082
SVM	0.954	0.954	0.953	0.954	0.2035
Naive Bayes	0.940	0.940	0.940	0.940	0.2378

Table 3 Confusion matrix of Random Forest with 10-fold evaluation

		Predicted Label	
		Personal users	Non-personal users
True Label	Personal users	916	9
	Non-personal users	39	213

Table 4 Confusion matrix of Decision Tree with 10-fold evaluation

		Predicted Label	
		Personal users	Non-personal users
True Label	Personal users	909	16
	Non-personal users	37	215

Tree is better and easier to understand, with explicit rules that can be easily understood and applied. The confusion matrix of the Random Forest classifier is provided in Table 3 and the confusion matrix for decision tree is provided in Table 4. Decision Tree and Random Forest are very close in performance, both achieving >95% accuracy. Decision tree tends to predict more non-personal users.

The decision tree result and rules are shown in Figure 3, which provides interesting information on the characteristics and differences between non-personal and personal users. The two major rules that can predict 1082 cases are:

- IF (humanFactorPer <= 1.4% and urlPerc > 90%) THEN **non-personal user**
- IF (humanFactorPer>1.4% and meanSpeed>3.16e-4 and repliedPerc<=84%) THEN **personal user**

This indicates that the sources for tweets (five apps) and embedded emotions in tweets (via sentiment analysis), which are combined in the *humanFactorPer* feature, are among the most important characteristics that can distinguish non-personal users.

Figure 4 shows the value distributions and differences between non-personal and personal users based on eight selected features, seven of which are used in the decision tree. First, it confirms that the lack of emotions or usage of location-based apps is a dominant characteristic of non-personal users, although there is also a large number of personal users who have similar feature values. Therefore, multiple features are needed to collectively distinguish non-personal users. Second, the two geographic features show very interesting patterns. The majority of non-personal users have frequent movements across county boundaries (indicated by the *cntyChangesPerM* feature). Examples of such non-personal users are those that use the same account to send tweets from different locations such as weather station alerts or advertisement bots that work from multiple locations. This is important to know, since tweets can be used for population mobility analyses and these non-personal tweets/users, if not removed, can affect the analysis result. Third, the *meanSpeed* feature also shows interesting patterns since most non-personal users have a relatively low


```
humanFactorPerc <= 0.014
|       urlPerc <= 0.9
|       |       speedLimitCntsPerM <= 1.2: L (26 / 5)
|       |       speedLimitCntsPerM  > 1.2:  N (2)
|       urlPerc > 0.9:  N (207 / 2)
humanFactorPer > 0.014
|       meanSpeed <= 3.16E-4
|       |       maxNumTwtsPerM <= 10: L (15 / 2)
|       |       maxNumTwtsPerM  > 10: N (11 / 1)
|       meanSpeed > 3.16E-4
|       |       repliedPerc <= 0.84:  L (875 / 20)
|       |       repliedPerc  > 0.84
|       |       |       normVar <= 0.79:       N (5 / 1)
|       |       |       normVar > 0.79
|       |       |       |       emotweetPerc <= 0.38: L (24)
|       |       |       |       emotweetPerc  > 0.38
|       |       |       |       |       friendIndicator <= 0.63: L (9 / 1)
|       |       |       |       |       friendIndicator  > 0.63: N (3)
```

Figure 3 Rules identified by the J48 Decision Tree. Prediction labels (*N* for non-personal users and *L* for legitimate users) are placed at the end of each tree leaf (corresponding to a rule). The numbers in a bracket mean (total users / wrong predictions)

speed on average but some non-personal users can also have a high speed. On the other hand, the speed distribution for personal users is a more normal distribution, with a mean value about 5 km/hr.

4 Evaluation and Reclassification

To evaluate the prediction power of the explicit rules shown in Figure 3, we compiled a different and new data set of tweets from the recent four-month period (October 7, 2013–February 8, 2014) for the 48 US contiguous states and Washington DC. This data set includes more than 300 million tweets and 4,109,330 unique users, among which 2,375,717 unique users have at least five tweets. We extract the same set of features for this new data set. The discovered rules in Figure 3 are applied to classify the 2,375,717 unique users The classification identified 61,555 non-personal users (2.6%).

To check the correctness of this classification, we randomly retrieved 100 predicted non-personal users and 100 predicted personal users. We then read the tweets of these users and manually assign a true label based on our judgment. Based on the manual labeling, the confusion matrix for the classification is constructed and shown in Table 5. Out of the 100 predicted spammers, 51 are true non-personal users and thus the precision is 51% and the recall rate 51/(51+2) = 96%. The overall accuracy is 74%. However, since the sampling is not random (i.e. we deliberately sampled more non-personal users disproportionally), these rates should be interpreted with caution. This reclassification shows that accuracy with the new data is not as high as with the training data. While further investigation is needed to improve the performance, it is already very useful considering that: (1) this is a new data set for a completely different time period; (2) spam activities may have evolved and are not well represented in the old training data set; and (3) non-spatial twitter spam detection does not achieve a much higher accuracy. Potential improvements can be made with a larger and more comprehensive training data set, which ideally should be dynamically updated to capture recent changes.

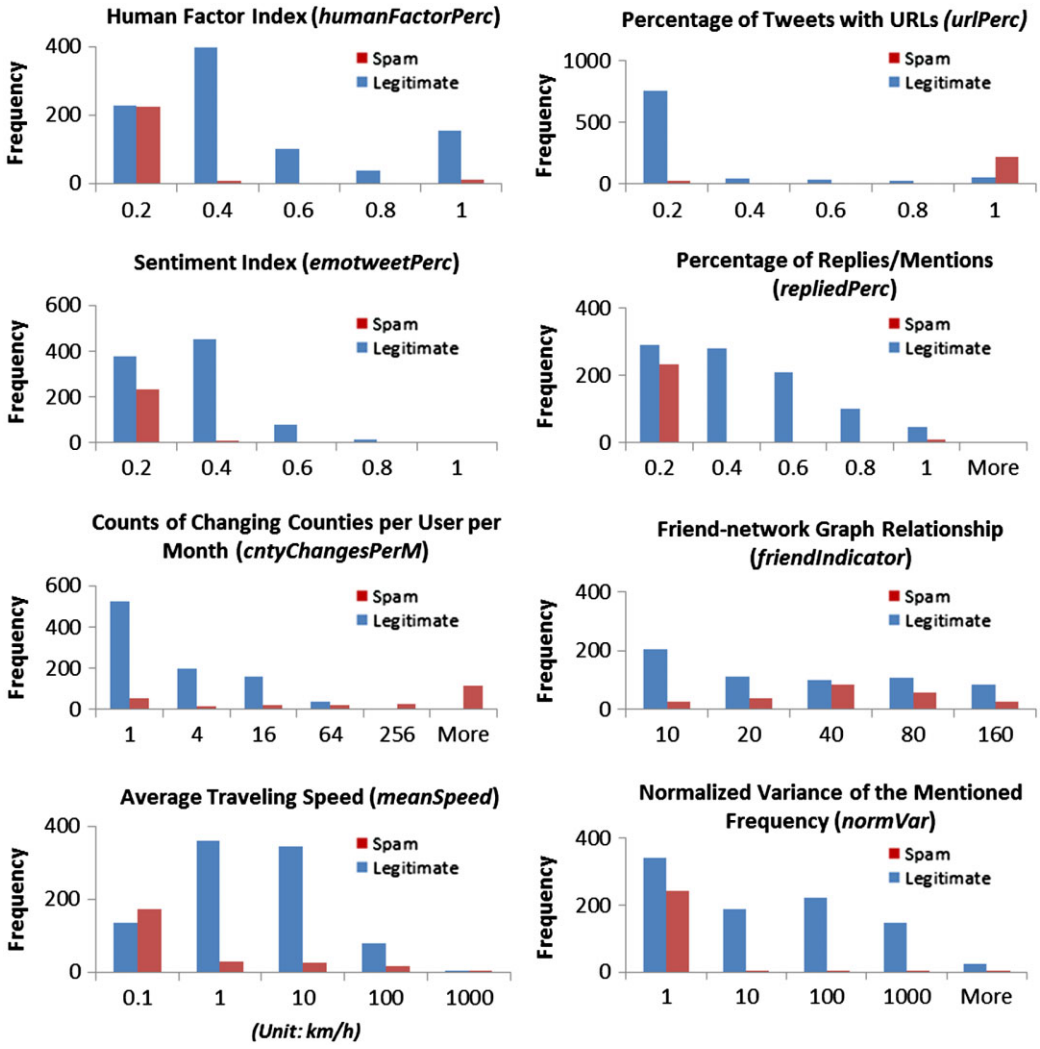

Figure 4 Comparison of non-personal users (labeled 'spam') and personal users (labeled 'legitimate') on eight selected features

Table 5 Reclassification confusion matrix with new data

		Prediction Label	
		Personal users	Non-personal users
True Label	Personal users	98	49
	Non-personal users	2	51

5 Discussion and Conclusions

In this article, we presented an analysis and detection approach for non-personal users on geo-tagged twitter network. With our definition, a non-personal user is one that does *not* post contents on personal daily lives, which can be a machine bot that automatically send tweets, an organization account that does not represent an individual, an individual that only sends non-personal information, or a spammer that posts commercial or malicious content. The focus of our analysis is on geo-tagged tweets with accurate locations (not just place names) and the users that have sent geo-tagged tweets. This research is important for understanding the characteristics of non-personal users on geo-social networks like Twitter and providing an effective means to detect and remove such users from subsequent analysis, in various applications which often assume that each twitter user is a real person. To the best of our knowledge, this is the first attempt to understand the behavior of non-personal users on a geo-tagged twitter network.

Our approach consists of multiple steps:

1. Extract user characteristics based on the geographic, graph-based and content information in tweets;
2. Construct training datasets by manually inspecting tweets and labeling a large sample of twitter users;
3. Conduct supervised classification and derive rules and knowledge for detecting non-personal users; and
4. Evaluate the derived rules with new training data and manual inspection. In this process, the design and extraction of effective features is very important.

We have extracted dozens of different features to describe the tweet contents, the friendship network, and the geographic patterns of each user. It remains a research question to extract more appropriate features for spam detection. Nevertheless, as a first attempt our study provides interesting and useful findings and insights about the differences between non-personal users and personal users according to their patterns. The top features that emerged in our study include not only known features such as the percentage of URLs and replies but also new features such as the mean tweeting speed, emotion, and location-based app usage.

We estimate that about 2–3% of twitter users with geo-tagged tweets are non-personal users. Many non-personal accounts involving geo-tagged tweets are normally limited to location-based services, such as job news, weather reports and local crime alerts. Some accounts that may be labeled as spammers in non-geotagged analysis are actually considered as personal users in our study, such as those who mainly post tweets from location-based apps like Foursquare. As our study shows, non-personal geotagged twitter users exhibit dramatically different patterns in terms of spatial mobility, such as frequent moves across county boundaries, different temporal patterns of movements, and different geo-social connections such as replies and mentions. Therefore, it is important to understand and filter out such outliers in a geo-social analysis of tweets or twitter users.

The classification accuracy with the training dataset is excellent, although the re-classification accuracy with the new data is not very high. There may be several reasons for this, which require further investigation. First, due to the time-consuming nature of the task, our training data set is still relatively small, with limited representation of the diversity and complexity of twitter activities. A larger and more comprehensive data set is needed, which is likely to improve the reclassification accuracy. Nevertheless, the discovered patterns have shown a clear difference between non-personal users and personal users. Second, non-personal

users, especially accounts operated by people, tend to demonstrate complex behaviors that are often a mix of non-personal and personal characteristics. On a fast changing social networking platform like Twitter, the spam activities also change quickly. To improve the detection schema, dynamic training data sets are needed that can capture and adapt to changes.

References

Benevenuto F, Magno G, Rodrigues T, and Almeida V 2010 Detecting spammers on Twitter. In *Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference*, Redmond, Washington

Bertrand De L, Robin S S, and Gianluca L 2009 "OMG, from here, I can see the flames!": A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the International Workshop on Location Based Social Networks*, Seattle, Washington

Breiman L 2001 Random forests. *Machine Learning* 45: 5–32

Castillo C, Donato D, Gionis A, Murdock V, and Silvestri F 2007 Know your neighbors: Web spam detection using the web topology. In *Proceedings of the Thirtieth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, The Netherlands

Chu Z, Gianvecchio S, Wang H, and Jajodia S 2012 Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing* 9: 811–24

Geng G-G, Li Q, and Zhang X 2009 Link based small sample learning for web spam detection. In *Proceedings of the Eighteenth International Conference on the World Wide Web*, Madrid, Spain

Go A, Bhayani R, and Huang L 2009 Twitter sentiment classification using distant supervision. *Processing* 150(12): 1–6

Goodchild M F 2007 Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research* 2: 24–32

Grier C, Thomas K, Paxson V, and Zhang M 2010 @spam: The underground on 140 characters or less. In *Proceedings of the Seventeenth ACM Conference on Computer and Communications Security*, Chicago, Illinois

Huberman B A, Romero D M, and Fang W 2009 Social networks that matter: Twitter under the microscope. *First Monday* 14(1): 1–5

Joachims T 1998 Text categorization with Support Vector Machines: Learning with many relevant features. In *Proceedings of the Tenth European Conference on Machine Learning (ECML '98)*, Chemnitz, Germany

MacEachren A M, Robinson A C, Jaiswal A, Pezanowski S, Savelyev A, Blanford J, and Mitra P 2011 Geo-Twitter analytics: Applications in crisis management. In *Proceedings of the Twenty-fifth International Cartographic Conference*, Paris, France

Martinez-Romo J and Araujo L 2013 Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems Applications* 40: 2992–3000

McCord M and Chuah M 2011 Spam detection on twitter using traditional classifiers. In *Proceedings of the Eighth International Conference on Autonomic and Trusted Computing*, Banff, Alberta

Mishne G 2005 Blocking blog spam with language model disagreement. In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '05)*, Chiba, Japan

Nigam K, Lafferty J, and McCallum A 1999 Using maximum entropy for text classification. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence Workshop on Machine Learning for Information Filtering*, Stockholm, Sweden

Ntoulas A, Najork M, Manasse M, and Fetterly D 2006 *Detecting spam web pages through content analysis. In Proceedings of the Fifteenth International Conference on the World Wide Web*, New York

Quinlan J R 1996 Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research* 4: 77–90

Sahami M, Dumais S, Heckerman D, and Horvitz E 1998 A Bayesian approach to filtering junk e-mail. In *Proceedings of the 1998 Learning for Text Categorization Workshop*, Madison, Wisconsin

Schneider K-M 2003 A comparison of event models for Naive Bayes anti-spam e-mail filtering. In *Proceedings of the Tenth conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary

Sitaram A and Bernardo A H 2010 Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Toronto, Ontario

Wang A H 2010a Detecting spam bots in online social networking sites: A machine learning approach. In Foresti S and Jajodia S (eds) *Data and Applications Security and Privacy XXIV*. Berlin, Springer Lecture Notes in Computer Science Vol. 6166: 335–42

Wang A H 2010b Don't follow me: Spam detection in Twitter. In *Proceedings of the International Conference on Security and Cryptography (SECRYPT 2010)*, Athens, Greece

Wu Y-S, Bagchi S, Singh N, and Wita R 2009 Spam detection in Voice-over-IP calls through semi-supervised clustering. In *Proceedings of the ACM/IEEE Conference on Dependable Systems and Networks*, Boston, Massachusetts: 307–16

Yang J and Counts S 2011 Predicting the speed, scale, and range of information diffusion in Twitter. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, Washington, DC

Yardi S, Romero D, Schoenebeck G, and Boyd D 2010 Detecting spam in a Twitter network. *First Monday* 15(1): 1–14

Zandbergen P A 2009 Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS* 13(s1): 5–25

Zhou D, Burges C J C, and Tao T 2007 Transductive link spam detection. In *Proceedings of the Third International Workshop on Adversarial Information Retrieval on the Web*, Banff, Alberta

