Spam Filtering in Twitter using Sender-Receiver Relationship

Group-V

Motivation

- Twitter is ranked in the top 10 most visited sites.
- In 2011, people sent about 140 million tweets per day and 460,000 new accounts were created per day.
- This popularity, however, also attracts spammers.
- Twitter spam is different.
- Twitter limits the length of each message to less than 140 characters.
- To overcome this restriction, spammers usually send a spam containing URLs that are created by URL shortening services.
- Since the messages are short and the actual spam content is located on external spam pages, it is difficult to apply traditional spam filtering methods based on text mining to Twitter spam.

Existing Methods

- Based on the characteristics of social networks.
- Honeypot-based approaches have been proposed . These studies created several honey-profiles and waited for spammers' contacts. They analysed the collected data and tried to automatically identify spammers by analysing spammer's behaviour.
- Based on statistical analysis .
- They also collected a large number of user profiles and manually classified the users into spammers and non-spammers.
- Finally they trained a classifier to identify spammers using data mining techniques.

Limitations

- First, account features such as tweeting interval, content similarity, age, the number of followings and the number of followers can be manipulated by spammers.
- Secondly, previous methods able to detect spammers only after spam has already been sent to legitimate users.
- To classify a user, previous methods need to know how a user has been tweeting and what a user has been tweeting.
- There is an inevitable delay between spam account creation and its detection.

Propose Method

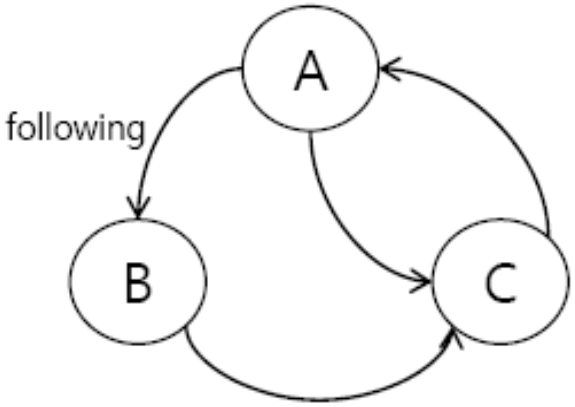
- Instead of account features, we considers the relation features between a message sender and a receiver, which are difficult for spammers to manipulate.
- Propose two relation features, which are distance and connectivity, to identify spammers. These relation features are unique features of social networks and are difficult for spammers to forge or manipulate.
- Propose system identifies spammers in real-time.

Twitter Features

- **Tweet-**In Twitter, both a post and posting action are called tweets.

ers.

-



; of

-

ion.

er

Fig. 1. Simple Twitter graph. User A is follower of user B and C and is also following of user C.

- **Retweet-** A retweet is a reposting another user's tweet.
- **Hashtag -**The '#' symbol is a hashtag in Twitter. The hashtag is attached to the front of keywords to categorize tweets.

How Twitter Deals with Spam

- Twitter has established several restrictions to prevent spam and abuse.
  - Following a large number of users in a short time
  - Following and unfollowing someone in a short time or repeatedly
  - A small number of followers compared to the amount of following
  - Multiple duplicated updates
  - Updates mainly consisting of links
- The above restrictions, however, are easy to avoid and spammers can always create new accounts even though their old accounts have been suspended.

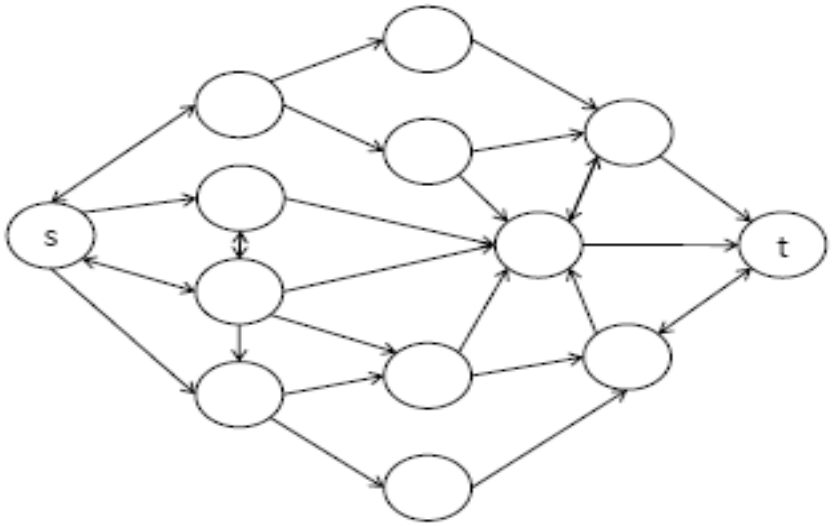
Overview

- First, we measure the **distance** of user pairs. For example, when two users are directly connected by a single edge, the distance between the users is one.
- In our experiment, almost all messages that come from a user whose distance is more than four are spam. Thus, the relationship is meaningless or untrustworthy when the distance is over four.
- The second feature is the **connectivity** between users which represents the strength of the relationships.
- An edge may exist between a legitimate user and a spammer when the spammer establishes a relationship with a legitimate user. i.e. called attack edges.
- Thus, the connectivity between a legitimate user and a spammer is weaker than the connectivity between legitimate users, when the

  distance is the same.
- We measure connectivity by using random walk and min-cut techniques.

Graph

- We construct the graph on the relation between the message sender
-



1

2

3

4

Fig. 2. A simple example of the graph when the distance is three.

- Analysing the relation between the receiver and the sender is the most important task in this work. We do not need an entire network .
- We use both the followings of the receiver and the followers of the sender to reduce crawling data. If we only use the receiver's followings, the amount of the crawling data will increase exponentially.
- We only analyse the user pairs whose distance is at least four. Moreover, Kwak et al. showed that 70.5% of user pairs have paths whose length is four or shorter in the Twitter network . Thus, our propose method covers most cases in Twitter.

Features

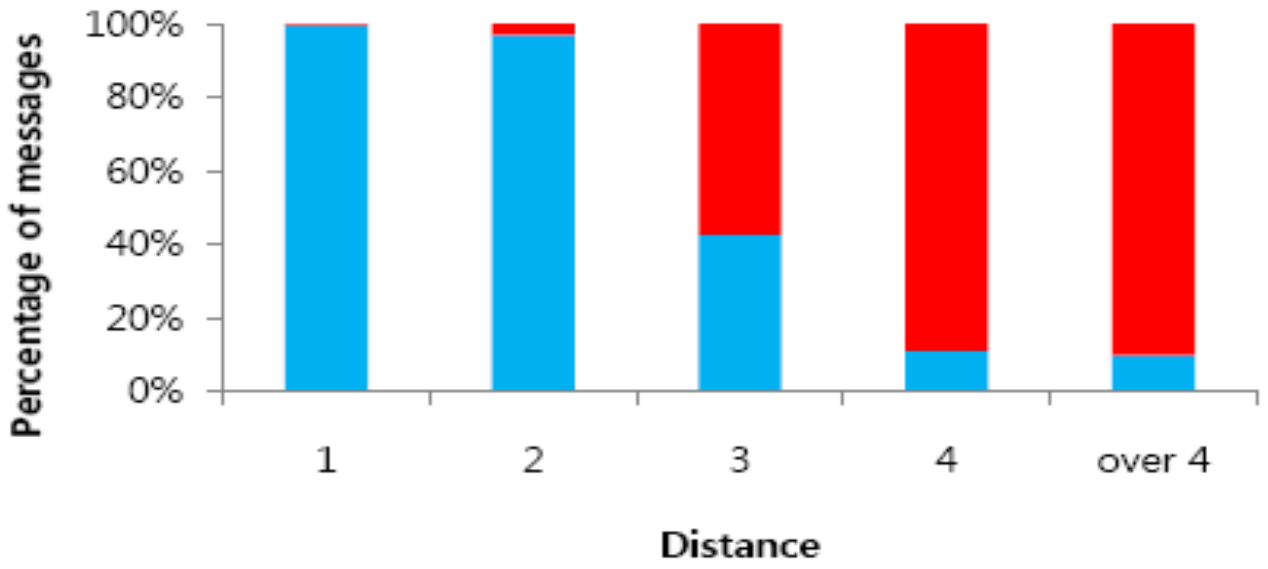


Fig. 3. The percentages of benign (blue) and spam messages (red).

- Menger's theorem defines edge-connectivity as follows:
- Let G be a finite undirected graph and u and v be two distinct nodes. The size of the minimum edge cut for u and v is the same as the maximum number of the edge-independent paths from u to v.
- As expected, the min-cut sizes of the spammer's cases are smaller than that of the normal cases.

Random Walk

- Random walk technique used in PageRank
- The idea behind PageRank is that when a random surfer visits pages infinitely, the pages linked more are visited more.
- PageRank values are computed by the left eigenvectors x_L of the transition probability matrix P
- $x_L P = \lambda x_L$,
- where λ is eigenvalue.

- The N entries in the eigenvector x_L are the steady-state probabilities of the random walk corresponding to the PageRank values of web pages.
- The Perron-Frobenius Theorem tell us that the largest eigenvalue of the matrix is equal to one which is the principal eigenvector.
- Thus, the principal eigenvector of the transition matrix P is the PageRank values.
- The web pages are corresponding to the users and the links are corresponding to the friendships.
- All edges point toward the node t. Thus the eigenvector of the node t is always top.
- Therefore, we convert the directed graph G' to the undirected graph G".
- All random walks will proceed to both nodes t and s in normal cases.
- When the node t is a spammer, however, the eigenvector of the node t will not be as high as the node s because the spammer only has a few edges.

Experiments and Evaluation

- **Data collection**

-Twitter offers API methods for data collection to encourage third-party developers, but there is a rate limit . A host is permitted 150 requests per hour.

-In order to overcome the rate limit we used four servers and

120 IP addresses. The servers changed their IP addresses when they were stopped by the rate limit.

-We randomly selected non-spammers by using numerical Twitter user IDs.

-Spam accounts were selected from among the reported accounts to the “@spam" account, which is the official Twitter account.

-Legitimate Twitter users can report the spam accounts by mentioning to the “@spam" account; thus, we searched mentions using the “@spam" keyword and collected spam accounts from the search results.

Spam Classification

- Results of random walk with the distance.
- Let i be the index of a receiver and j be the index of a sender in x_L. Then, their random walk values are $x_L[i]$ and $x_L[j]$, respectively.
- When the sender is a non-spammer, $x_L[i]$ and $x_L[j]$ are similar values and they are quite higher.
- When the sender is a spammer, $x_L[j]$ is much lower than $x_L[i]$.
- We use the ratio $x_L[j]=x_L[i]$ as a feature from random walk.
- We used Weka ,which is a data mining tool, and used 10-fold cross validation option in classification.

- A
- T
- F

Table 1. The results of classification using distance and random walk

Classifiers	True Positive (%)	False Positive (%)
Bagging	93.3	8.5
LibSVM	93.2	8.3
FT	93.1	7.7
J48	92.3	8.7
BayesNet	92.0	8.0

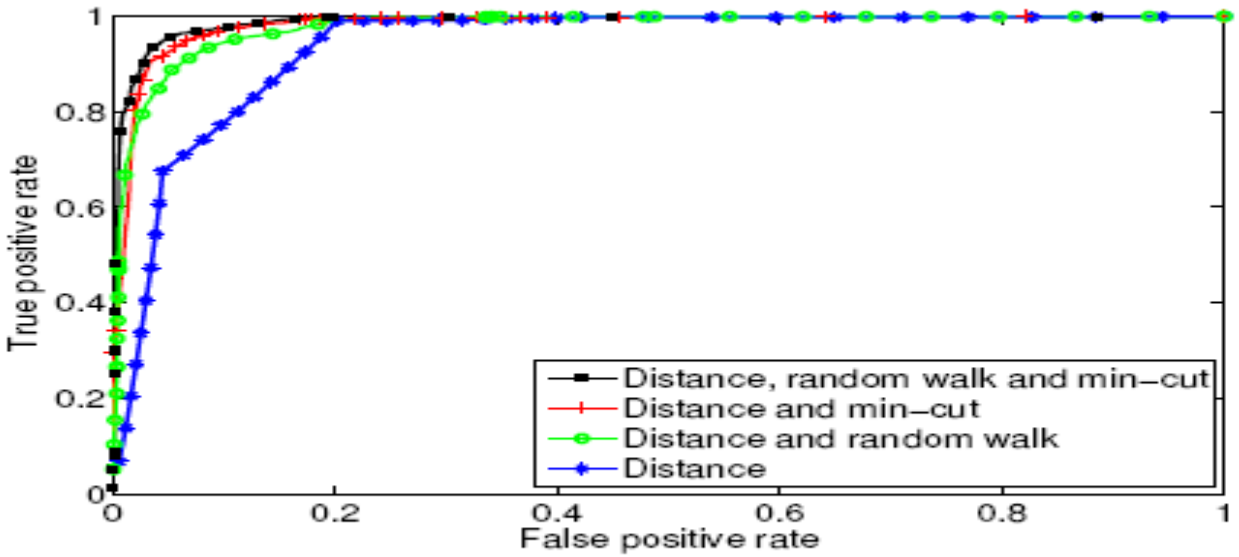


Fig. 5. ROC curves for each of the relation features.

Spam account detection with including a user relation feature

- The 11 features that are used in classications are as follows:

- The standard deviation of tweeting interval
- The ratio of tweets containing URLs
- The ratio of mentions containing URLs
- The ratio of tweets containing hashtags
- The ratio of mentions (|mentions|/|total tweets|)
- The ratio of duplicate tweets
- Reputation (|followings|/|followers|)
- The number of lists including the user
- Age (the current time - the account creation time)
- The average content similarity
- The ratio of mentions sent to non-followers

- The ratio of mentions sent to non-followers is the only relation feature

Table 4. The ; Weka classifiers

C	(%)
B	
L	
J4	
L	
Li	

Table 5. The results of feature selection

Rank	Information Gain
1	The ratio of mentions sent to non-followers
2	Reputation
3	The ratio of mentions containing URLs
4	The ratio of tweets containing URLs
5	Age

Rank	ReliefF
1	The ratio of mentions sent to non-followers
2	The ratio of tweets containing URLs
3	Age
4	The ratio of mentions containing URLs
5	The average content similarity

Rank	Chi Square
1	The ratio of mentions sent to non-followers
2	Reputation
3	The ratio of mentions containing URLs
4	The ratio of tweets containing URLs
5	Age

- If we use both the account features and the relation features, the spam filtering system will be more powerful.
- The accuracy is better than the results when only used the relation features.
- Our system can be applied to both client-side and server-side.
- There will not be resource and computation problem at client.
- Additional bandwidth and storage resources are not needed at server-side because service managers already have user's relation information.
- Computed relation features will be cached and then only updated when the relation features are changed in order to reduce computing overhead.

Limitations

- First, if a normal user creates a new account and sends a message to his friend before the new account has any followers, the message will be filtered.
- This, however, is a temporal problem because the new account will get followers soon.
- The second problem is that our system will identify the messages as normal even though the messages come from infected friends.
- When a user sends the messages using the application that has never been used by the user, the messages should be suspected. Ultimately, the contents of the messages should be checked whether the contents are spam or not.