Predicting Individual’s Posting Behaviors in Social Network with Markov Chain

Fengcai Qiao¹, Su He² and Jinsheng Deng¹

¹College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, China
²China Aerospace Science and Technology Corporation, Shenzhen, China
Email: fcqiao@nudt.edu.cn

Abstract. Online social networks like twitter are self-organized systems with emergent behaviors from the individual interactions. Predicting the users’ behaviors like posting, replying or retweeting is one of the most important issue. In the paper, we present a method to model the twitter users’ posting behavior with the other users’ response. We also contrast the simulation precision with the model that does not involve the other users’ response. We make use of a stochastic model with Markov chain to predict a specific user’s next tweeting behavior. From the result by contrasting the simulation data and the real data, we demonstrate that the engineering method is able to predict individual posting behaviors based on time windows. In addition, the proposed model involving other users’ response is better than that only considering posting time sequences.

1. Introduction
Twitter has been broadly applied to numerous fields as a tool for receiving immense news, communicating with friends, sharing information, keeping up with favourite teams, expressing own point of view and emotions, marketing, participating in politics and so on. Different kinds of people use it for different purposes. These activities generate huge data about people’s tweeting, replying and retweeting behaviors that provide probabilities for learning models of user’s behavior and interests. Personality tweeting behavior on twitter is an important trait that reflects people’s online habits, interests and interactions with other people. In order to analyze this behavior, we must know factors that impact people’s tweeting behavior. Therefore, a fundamental set of questions is following: When will user post tweets? Why user post tweets? What factors affect user’s tweeting behavior (own interests, other users’ replies or retweets)? Answers to these questions can help us to gain an insight into the mechanism of personality tweeting behaviors.

In this paper, we consider a huge twitter dataset, consisting of all tweets published by the users and their followers, replies and retweets about their published tweets. We focus on the analyzing of the influence of posting behavior, viewing behavior and other people’s replying and retweeting behavior to one user’s next tweeting behavior. Firstly, when a user login into Twitter, it typically will see the timeline stream of tweets sent by all of his followers. When an interested tweet comes into view, he may retweet it or post his own opinion about it. Secondly, if his tweets are retweeted or replied by others, it can encourage the user to publish more tweets next time, which is called self-confidence.

Researchers have developed a series of probabilistic models to learn user behaviors [1][2][3]. But they only considered users’ tweeting behavior, ignoring the influence of interaction with others. As we analyzed above, the user’s viewing content and other people’s participation are important factors that affect one user’s next tweeting behavior. As a result, we introduce stochastic model to build users’
tweeting behavior. Stochastic modelling is an important tool for modelling user behavior. It is a probabilistic framework that represents each user as a stochastic process with different states. Rather than using fixed variables as in other mathematical modelling, stochastic model incorporates random variations to predict future conditions, which is suitable for predicting user behavior.

We proposed a stochastic model using k-dependent Markov chain to predict the user’s next tweeting behavior. We consider not only the specific user’s tweeting behavior but also other users’ retweet and reply behavior. We think that retweet and reply behavior from others can increase the sense of achievement and then encourage the user to perform more active.

The paper is organized as follows. The data used is provided in Section 2. The predicting model is presented in Section 3. In Section 4, the simulation experiments give the results and analysis. The work is summarized and conclusions drawn in Section 5.

2. Dataset

2.1 Data Collection

The data used in this paper consist Twitter posts related to initiatives about American winter storm. We collected the data using the following strategy. First, we collected a series of hash tags related to the topic of American winter storm, such as “#storm”, “#winter storm”, “#American storm” and so on. Next, we used these hash tags to search tweets and user’s id. After getting users’ id list, we collected all their tweets including tweet content, publish time, reply number, retweet number and favourite number. Finally, we gathered each user’s followers and their tweet information. The total collected flow is shown as Figure 1.

We used this strategy to monitor Twitter for two weeks. We totally collected 9000 users whose tweets contain the hash tags. The totally dataset consists of 1 million users and more than 750 million tweets.

2.2 Data Overview

In order to learn more about the data, we make a brief analysis about the posting number distribution, the posting interval distribution and the online time distribution.

2.2.1 Posting Number Distribution. Based on the time window, which is set to 30 minutes in this paper, each user’s posting time sequence was divided into a series of time windows with posting number in it, as shown in Figure 2. We can know that most of the posting number is less than five.
2.2.2 Posting Interval Distribution. After dividing the users’ posting time sequences into time windows and the statistical analysis of consecutive time windows, we see that most of the posting number is nearly zero. Figure 3 illustrates that the posting interval distribution takes on power-law characteristics. That is to say most users post *tweet* in a fixed time.

![Figure 3. Posting interval distribution of the dataset](image)

2.2.3 Online Time Distribution. We identified an online user by his actions, including posting, retweeting, replying. If a user doesn’t do these actions in five minutes, we set his status to offline. From Figure 4, we know that most users’ online time is less than 30 minutes.

![Figure 4. Users’ online time distribution of the dataset](image)

3. Model

3.1 Base Idea

The behaviors of individuals on twitter depend on the users’ habits, which can be expressed through the history of posting, viewing and interactions with other users. We consider that the posting behavior, viewing behavior and other user’s responding behavior can be used for predicting one user’s next posting behavior. This prediction process has Markov property [4]. A Markov process is succinctly captured by a state diagram showing the possible states and transitions between these states. Developing a model for posting behavior requires identifying the main states and relevant transitions between these states. Each state represents the probability of performing the specified action, with transitions leading to conditional probabilities.

3.2 Prediction Model

In this section, we focus on the detail of our proposed model. There are several actions in twitter that we can observe in the data including posting, viewing, favourite, retweeting or replying. In our study, we make favourite, retweeting and replying behaviors as a unit, calling *responding*, which represents
the other user’s response to a message. To learn this behavior, the model receives the list of the users in twitter as input. Each user’s data containing his or her posting timelines, comments, and the response from other users. From these data, the users’ state change transitions are modelled as a Markov chain where the current state depends on the previous state. The following assumptions are considered in our model:

- The posting time is discrete and we consider a ∆t time interval to define action time windows;
- The user’s actions like posting, comments and responding are computed on these time windows.

In addition, the Markov states are also attached to the time windows. Therefore, the current state means that the user posted in the current time window, while the previous state means the user posted, his or her following users posted and other users replied or retweeted in the previous time window.

We calculate the number of the user’s own tweets, the number of the user’s timeline tweets and the number of other users’ response for each time window. Then we get a list of time windows, each time window contains three numbers: the number of posting, the number of viewing and the number of response. Given the history data, our work focus on the prediction of the three numbers in the next time window. The totally process flow is shown in Figure 5.

![Figure 5. The flow of the prediction model.](image)

Denote \( W_{t-1}, V_{t-1} \) and \( R_{t-1} \) be the posting, viewing and responding vector at time \( t - 1 \). Figure 6 describes the transitions and states that can be observed in the data. Empty vectors \((V = \phi, W = \phi, R = \phi)\) mean non-observed data. State \( S_4, S_5, S_{12} \) and \( S_{13} \) are impossible to reach, because if the users do not post tweet in a time window, other users definitely can’t respond to the tweets in the same time window.

![Figure 6. Transitions and states that can be observed](image)

We calculate the Maximum Likelihood Estimation, denoted as \( L \), to determine which transition will be reached in the next time window. Therefore, for each user, we estimate \( L \) as follows:

- Observed transitions:
  \[
  L(S|V_{t-1}, W_{t-1}, R_{t-1}, W_t) = \frac{\text{count}(S, V_{t-1}, W_{t-1}, R_{t-1}, W_t) + 1}{\text{count}(S, V_{t-1}, W_{t-1}, R_{t-1}, W_t) + |S|}
  \]

- Unobserved transitions:
  \[
  L(S|V_{t-1}, W_{t-1}, R_{t-1}, W_t) = \frac{1}{\text{count}(S, V_{t-1}, W_{t-1}, R_{t-1}, W_t) + |S|}
  \]
Where \(|S|\) denotes the number of states. After calculating the transition probability for all states, we get transition probability matrix \(P\), where \(n=|S|\).

\[
P = \begin{bmatrix}
p_{11} & p_{12} & \cdots & p_{1n} \\
p_{21} & p_{22} & \cdots & p_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
p_{n1} & p_{n2} & \cdots & p_{nn}
\end{bmatrix}
\]

From transition probability matrix \(P\), we choose \(\text{Max}(P)\) as the next transition that will happen. Then we can know whether the user will post a tweet or not. After that, we need to estimate three parameters for the next time window: tweeting number, viewing number and response number.

### 3.3 Parameter Estimation

#### 3.3.1 Posting Number
If the transition state in the next time window is not posting, then \(W_t = 0\). Otherwise, we calculate the posting number from history tweeting data. We selected recent \(N\) time windows but the time window whose posting number is zero. Then we chose the median of these posting numbers from the \(N\) time windows. If all the posting numbers in the \(N\) time windows are zero, we set \(W_t = 1\).

#### 3.3.2 Viewing number
The user’s viewing number is related to his followed users’ posting behavior, and all the tweets of his followed users’ tweets are shown in the user’s home page. We chose the recent \(N\) time windows to predict viewing number in the next time window. As the posting number, the median of the viewing numbers from these \(N\) time windows is chosen as the final viewing number.

#### 3.3.3 Response number
Responding behavior is more complex than posting and viewing behavior, which is consisted by three parts: replying, retweeting and favourite. Just as predicting posting and viewing number, firstly we selected recent \(N\) time windows as history data, then we chose the median of the responding numbers from these \(N\) time windows.

### 4. Experiments

#### 4.1 Setting
For each user, we use history data to predict his future behavior. The user’s behavior is determined by the Markov Chain Monte Carlo simulation method.

Before simulation, two important steps were conducted:

i) We divided the posting time sequence to a series of time windows (30 minutes as a window);

ii) We chose previous half windows as initialized data.

We do the simulation in two different scenarios (involving other users’ responding and non-involving other users’ responses) and contrast the results of these two scenarios.

#### 4.2 Validation
The Root Mean Square Error (RMSE) was frequently used to validate the models for evaluating the difference between two time series. The formula to calculate as equation:

\[
RMSE(T) = \sqrt{\frac{1}{T} \sum_{i=1}^{T} (y'_i - y_i)^2}
\]

In this equation, \(y'_i\) denotes the predicted number of tweets posted at time \(t\), and \(y_i\) represents the actual number of tweets posted at time \(t\) in the observed data.

Our proposed model is validated using the Coefficient of Variation of the Root Mean Square Error (\(CV_{RMSE}\)). The results of the prediction model are compared with those computed from the observed data.
data. The $CV_{RMSE}$ normalizes to the mean of the observed data. Hence, using this metrics we can compare both pattern and volume.

$$CV_{RMSE} = \frac{RMSE(T)}{Avg(y)}$$

Where Avg(y) represents the average number of tweets posted in the time T.

4.3 Results
In this section we present the experiments results. The main purpose is to compare the predicted number of posted tweets with the real number of tweets posted by the user in the observed data.

For each user we ran 10 simulation trials and computed the average. For each user we predicted tweeting number for a thousand of time windows, and calculated the precision of simulation result and the real data. As shown in Figure 7, although the tweeting number in the simulation data can’t equal to the real data for each time window, but the judgement of whether a user posts tweet or not for each time window is relatively accurate. Actually, the accuracy of posting achieves 80%.

Figure 7. Comparison of the simulation and the real data.

In addition, in our experiment, we contrasted two scenarios: involving users’ response (response) and only involving users’ posting behavior (non-response). For both scenarios, we calculated the $CV_{RMSE}$ of the simulation sequences and the real data sequences. From Figure 9, it can be observed that the error rate of the prediction model involving response is generally lower than that only considering users’ posting behavior. This indicates that the users’ response can improve the overall accuracy of the prediction model, which is consistent with our assumption above: other users’ response can motivate a user’s posting behavior.

5. Conclusion
In this paper we proposed a method to simulate the behavior of users in twitter, using the data we collected from twitter about American winter storm. The data sampled from twitter allowed us to build
individual stochastic models to represent how each user behaves when posting messages and how the users interact with each other in the social network. Experiments considering two different scenarios demonstrated that the proposed approach is promising for simulating the overall behavior of users in twitter, and the parameter of response behaviors improved the accuracy of simulation result.

From the proposed method, the future work is two folds. First, we need to enlarge the dataset, in that larger dataset may help us better estimate the model. In our next step of experiment, we plan to do our simulation based on Spark [5], enhancing our calculating capability. Second, we need to enhance the modelling of users. This may be achieved by using the machine learning and optimization techniques, involving the topic of tweets, sentiment, interests and the relation of users.

6. Acknowledgment
This work was supported by the Scientific Research Project of National University of Defense Technology (No. JS19-01).

7. References
[1] J.L. Iribarren and E. Moro, “Impact of human activity patterns on the dynamics of information diffusion,” Physical Review Letters, vol. 103, 2009.
[2] K. Lerman and T. Hogg, “Using stochastic models to describe and predict social dynamics of web users,” ACM Transactions on Intelligent Systems and Technology, vol. 3, no. 4, September 2012.
[3] I Zhao, Peihai, et al. "Behavior Analysis for Electronic Commerce Trading Systems: A Survey." IEEE Access 7 (2019): 108703-108728.
[4] Grewal, Jasleen K., Martin Krzywinski, and Naomi Altman. "Markov models—Markov chains." (2019): 663-664.
[5] Zaharia, Matei, et al. "Spark: Cluster computing with working sets." HotCloud 10.10-10 (2010): 95.