# Predicting Retweeting Behavior Based on Autoregressive Moving Average Model

Zhilin Luo[1,2], Yue Wang[2], and Xintao Wu[2]

[1] Northwestern Polytechnic University, China
[2] University of North Carolina at Charlotte, USA
{zluo5,ywang91,xwu}@uncc.edu

**Abstract.** In this paper, we consider a fundamental social network issue that illustrates how information dynamically flows through a social media network. Inferring the number of times that a particular message posted by some specific user will be retweeted by his followers and predicting the number of readings of the posted message via various retweeting chains are central to understanding the underlying mechanism of the retweeting behaviors. Specifically we work on the Task 2 of the WISE 2012 Challenge, i.e., predicting retweet behaviors in the Sina Webo data set. We develop an approach based on the Autoregressive-Moving-Average (ARMA). In the approach, we treat retweeting activities of each original tweet as a time series where each value corresponds to the number of times that the original tweet is tweeted or the number of times of possible-view of the original tweet during that particular time period. For each tweet in the test data, our approach first identifies the most similar message from the training data based on the similarity between their time series values in the same length period as provided in the test tweet, fits the ARMA models over the whole time series of the identified message, and then applies the fitted model over the time series of the test tweet to predict future values. We report our prediction results and findings in this paper.

## 1 Introduction

Retweeting is one of the most important features in microblogging sites such as Twitter and Sina Weibo and examining retweeting behavior has been an active research area recently [5–8]. The retweeting mechanism empowers users to spread their ideas beyond the reach of the original tweet's followers. The process can be regarded as information diffusion in social media.

In this paper, we consider a fundamental social network issue that illustrates how information dynamically flows through a social media network. Inferring the number of times that a particular message posted by some specific user will be retweeted by his followers and predicting the number of readings of the posted message via various retweeting chains are central to understanding the underlying mechanism of the retweeting behaviors. In Microblogging, some particular messages posted by particular users are often retweeted widely and promptly

X.S. Wang et al. (Eds.): WISE 2012, LNCS 7651, pp. 777–782, 2012.

while others attract little attention from other users. Various factors including the message content, associate event, the timing, and the local structure of the followship network may influence the message propagation.

Specifically we work on the Task 2 of the WISE 2012 Challenge [1]. WISE 2012 Challenge is based on a data set collected from Sina Weibo, one of the most popular Microblogging service. The followship network is also provided. In the test data set, a small part of retweeting activities of thirty-three tweets of six events are given. In the challenge, we are required to predict the retweeting activities of those thirty-three tweets. Specifically, we are required to predict two measurements of the original tweet after 30 days:

1. M1: the number of times that the original tweet is retweeted.
2. M2: the number of times of possible-view of the original tweet. The number of possible-view of a tweet is defined as the sum of all possible-view numbers of retweet actions.

In this paper, we study the use of Autoregressive-Moving-Average(ARMA) models to predict retweeting behaviors. ARMA models are mathematical models of the persistence, or autocorrelation, in a time series. ARMA modeling is effective to understanding the physical system by revealing the physical process that builds persistence into the series and predicting the behavior of a time series from past values alone. We treat retweeting activities of each original tweet as a time series where each value corresponds to the number of times that the original tweet is tweeted (for M1) or the number of times of possible-view of the original tweet (for M2) during that particular time period. For each tweet in the test data, we first identify the most similar message from the training data based on the similarity between their time series values in the same length period as provided in the test tweet, fit the ARMA models over the whole time series of the identified message, and then apply the fitted models over the (short) time series of the test tweet to predict future values.

## 2 Predicting Retweeting Behavior

Autoregressive-moving-average(ARMA) models, also called Box-Jenkins models, are mathematical models of the persistence, or autocorrelation, in a time series. We will work with the mean-adjusted series

$$y_t = Y_t - \overline{Y}, t = 1, ..., N \tag{1}$$

where $Y_t$ is the original time series, $\overline{Y}$ is the sample mean, and $y_t$ is the mean-adjusted series.

The autoregressive model includes lagged terms on the time series itself, and the moving average model includes lagged terms on the residual. We acquire the ARMA model by including both types of lagged terms.

[1] `http://www.wise2012.cs.ucy.ac.cy/challenge.html`

**Definition 1.** *(Autoregressive-moving-average model [2]) The $ARMA(p,q)$ refers to the model with p autoregressive terms and q moving-average terms. This model contains the $AR(p)$ and $MA(q)$ models,*

$$y_t + \sum_{i=1..p} a_i y_{t-i} = e_t + \sum_{i=1..q} c_i e_{t-i} \tag{2}$$

*where $a_1, ..., a_p$ are the autoregressive coefficients; $c_1, ..., c_q$ are the first-order,..., qth-order moving average coefficients; and $e_t, ..., e_{t-q}$ are the regression residuals at times $t, ..., t-q$.*

Recall we have two prediction measurements: M1 (the number of times that the original tweet is retweeted) and M2 (the number of times of possible view of the original tweet). Retweeting activities of each original tweet are modeled as a time series. Specifically, for M1, we treat the number of retweeting times of a message as $\mathbf{y} = \{y_1, y_2, \cdots, y_N\}$, where $y_t (t = 1, \cdots, N)$ denotes the number of times that the original tweet is tweeted at the time stamp $t$. $\mathbf{y}$ can be extracted from the retweet traces. We leave detailed discussions on how to transform tweet traces to time series in Section 3.1. We use $\mathcal{Y}$ to denote the set of time series from all original tweets in the training data. Similarly we use $\mathcal{Z}$ to denote the set of time series from all original tweets in the test data. The retweeting activities of each tweet in the test data are modeled as $\mathbf{z} = \{z_1, \cdots, z_s, z_{s+1}, \cdots, z_N\}$ where $z_t (t = 1, \cdots, s)$ denotes the observed number of retweet times at the time stamp $t$ and $z_t (t = s+1, \cdots, N)$ corresponds to the unknown number of future retweet times. Similarly, for M2, the time series can be extracted from the retweet traces and the followship network. The number of possible-view activity is defined as the number of followers of the user who conduct the retweet action. Each value in the time series denotes the number of possible-view at a particular time stamp.

**Algorithm 1.** ARMA based Prediction

**Input:** Training data $\mathcal{Y} = \{\mathbf{y_1}, \cdots, \mathbf{y_n}\}$
Tweet under test $\mathbf{z} = \{z_1, \cdots, z_s, z_{s+1}, \cdots, z_N\}$ where $z_{s+1}, \cdots, z_N$ are unknown
**Output:** $\sum_{t=1}^{N} z_t$

1: Identify $\mathbf{y} \in \mathcal{Y}$ such that $\mathbf{y}$ and $\mathbf{z}$ are from the same user and $\sum_{t=1}^{s} (y_t - z_t)^2$ is minimized;
2: Fit ARMA(p,q) over $\mathbf{y} = \{y_1, \cdots, y_N\}$;
3: Apply the fitted ARMA(p,q) over $\mathbf{z}$ to generate $z_t$ $(t = s+1, \cdots, N)$;
4: Output $\sum_{t=1}^{N} z_t$.

Algorithm 1 shows our approach. For each testing message, we firstly identify the most similar message in the training data set by minimizing the Euclidean distance of the partial time series among candidate messages that are authored by the same user. Secondly, we use the whole time series of the identified message to build the ARMA model. Thirdly, we apply the fitted ARMA model on the test tweet to generate future values of the time series. Finally we acquire the prediction results.

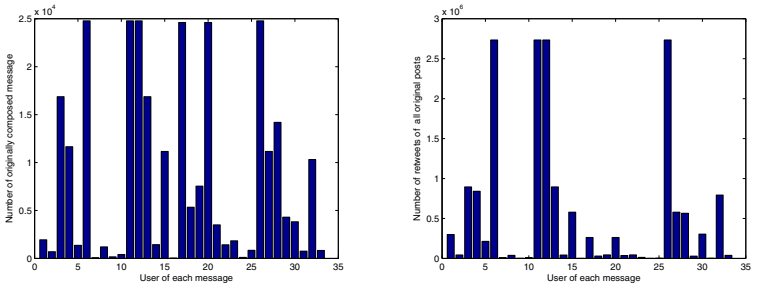
ARMA modeling proceeds by a series of well-defined steps. The first step is to identify the model, which consists of determining the structure(AR,MA or ARMA) and the order of the model($p$ and $q$). The second step is to estimate the coefficients of the model [3]. The third step is to check the model, which ensures the residuals of the model are random and the estimated parameters are statistically significant. In the paper, we use the $'Forecasting'$ module from SPSS package for ARMA model fitting and forecasting. We use R-square and root-mean-square error (RMSE) to choose parameters $p$ and $q$.

## 3 Evaluation

### 3.1 Data Preprocessing

The data set contains two types of data: tweets and followship network. Tweets includes basic information about tweets (time, user ID, message ID etc.), mentions (user IDs appearing in tweets), retweet paths, and whether containing links. In our prediction, we do not use information on whether containing links or topic information. The test data file contains retweets of thirty-three original messages composed by twenty-seven users. The thirty-three messages are from six events. The number of retweets for each original message in the test data is around 100. The retweet time period T for each original message can be easily derived from the field *rtTime* and the field *time* of the message's last retweet. We observe that the retweet periods of some messages are very short (less than 1 minute) whereas the retweet periods of other messages are quite long (more than 1 day). We then divide each time period T to $s$ bins, and derive the time series $\mathbf{z}$ where each value $z_t$ $(t = 1, \cdots, s)$ corresponds to the number of retweets at the time point $t$. We search the training data set to get candidate messages authored by the same user. For each candidate message, we follow the same strategy to derive its time series. We then identify the candidate message with time series $\mathbf{y}$ for ARMA model fitting such that $\sum_{t=1}^{s} (y_t - z_t)^2$ is minimized. We then run the fitted ARMA model over the test data $\mathbf{z}$ to derive future prediction values $z_t$ $(t = s+1, \cdots, N)$ where $N$ corresponds to the time period of 30 days. Finally, we output $\sum_{t=1}^{N} z_t$ as the predicted number of times that the message is retweeted in 30 days.

To predict the number of possible-view (for M2), we need to incorporate the number of followers for users who retweet the message. We can easily derive this information from the followship network. The thirty-three messages are composed by twenty-seven authors. There are four users who are not contained the followship network provided in the challenge. We show the distribution of the number of tweets authored by these twenty-seven users in Figure 1(a) and the distribution of the number of total retweets of messages authored by each of these twenty-seven users in Figure 1(b). From the figures, we can observe the distribution of the number of composed tweets and the distribution of the number of times of posts being retweeted are unevenly distributed, which causes the difficulty of prediction tasks.

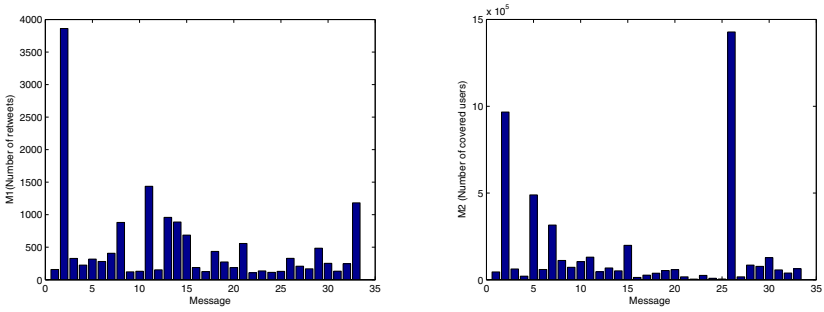



(a) Number of composed tweets (b) Total number of times of posts being retweeted

**Fig. 1.** Statistics of twenty-seven users in the test data

### 3.2 Prediction Results

Figure 2 shows the prediction result respectively for M1: the number of retweeted times of each message after 30 days of original post, and M2: the number of covered users of each message after 30 days of original post. For M1, the thirty-three test tweets received on average 487 times of retweets in the period of 30 days. The second message from the event of *Death of Steve Jobs* received the largest number of retweets (3862) whereas the second message from the event of *Xiaomi Release* received the smallest number of retweets (107). For M2, the thirty-three test tweets received on average 147,831 times of possible-view in the period of 30 days. The first message from the event of *Yao Jiaxin Murder Case* received the largest number of possible-view (1,428,387) whereas the last message from the event of *Xiaomi Release* received the smallest number of possible-view (4,200).




(a) M1 result (b) M2 result

**Fig. 2.** Prediction results

## 4 Conclusion and Future Work

In this paper, we introduced our approach based on ARMA models to predict retweeting behaviors in WISE 2012 Challenge. We expect that ARMA models can be used to help understand the underlying mechanism of retweeting behaviors. There are some other aspects of this work that merit further research. Among them, we will continue the line of this research by incorporating various factors (e.g., event information associated with a tweet and the fine-grained local topological structure of the followship network) in the modeling process. We will also explore advanced fitting strategies of ARMA models. In analyzing ARMA time series, it is typically assumed that only one realization is available for model fitting. In our current approach, for a given test tweet, we identify the most similar message to build the ARMA models. It is preferable to use multiple similar messages (from the same author or with similar topics) for model fitting. One strategy is that the elements of each time series are averaged across each time point to produce one series for analysis. In [1], the authors show that multiple independent time series from the same ARMA process can be represented by a single univariate ARMA time series through an interleaving of the original series. We will explore these strategies in our future work.

**Acknowledgments.** This work was supported in part by U.S. National Science Foundation IIS-0546027, CNS-0831204, CCF-0915059, and CCF-1047621. We would like to thank Sina Weibo to provide the retweet data set to the research community and thank the WISE organizers to provide this opportunity for us to participate this challenge. For detailed results, please refer to [4].

## References

1. Bowden, R.S., Clarke, B.R.: A single series representation of multiple independent arma processes. Journal of Time Series Analysis 33(2), 304–311 (2012)
2. Brockwell, P., Davis, R.: Time series: theory and methods. Springer (2009)
3. Chatfield, C.: The analysis of time series: an introduction, vol. 59. CRC Press (2004)
4. Luo, Z., Wang, Y., Wu, X.: Predicting retweeting behavior based on autogressive moving average model. Technical Report, UNC Charlotte (2012)
5. Luo, Z., Wu, X., Cai, W., Peng, D.: Examining multi-factor interactions in microblogging based on log-linear modeling. In: ASONAM (2012)
6. Macskassy, S.A., Michelson, M.: Why do people retweet? anti-homophily wins the day!. In: ICWSM (2011)
7. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: WWW (2011)
8. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: CIKM (2010)