Journal of Computers Vol. 31 No. 4, 2020, pp. 273-287
doi:10.3966/199115992020083104021

A Novel Method for the Time Series Data Processing and Analysis of Social Networks

Guixun Luo[1], Zhiyuan Zhang[2*], Zemin Bao[3], Naiyue Chen[1]

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
{98940279, nychen}@bjtu.edu.cn

[2] School of Electronic and Information Engineering, Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, Beijing Jiaotong University, Beijing, China
zhangzhiyuan@bjtu.edu.cn

[3] National Computer network Emergency Response technical Team/Coordination Center of China, Beijing, China
baozemin@cert.org.cn

Received 3 April 2020; Revised 3 May 2020; Accepted 3 June 2020

Abstract. Predicting the characteristics of weibo reposts through time series analysis is very important topic in the research area of social network. An improved method by introducing the wavelet transform mechanism based on the ARMIA model is used in this paper. We propose an improved data processing method based on wavelet theory to remove noise in time series and to enhance the ability of predicting the weibo reposts. Experimental results show that multi-scale analysis combined with high-frequency zeroing can retain the effective part of the data and remove the noise part of the high-frequency, making the curve smooth.

Keywords: socila network, time series analysis

1 Introduction

The information of weibo can be transmitted continuously through the repost function [1-2]. The size of the information repost scale is an important index to measure the effect information transmission. Therefore, predicting the characteristics of weibo reposts (e.g. future repost number of the target weibo, rules of weibo repost, time series characteristics of weibo repost) through time series analysis is very important topic in the research area of social network [3].

The prediction of weibo repost can be divided into two aspects [4]: sub-bureau prediction and global prediction. From the perspective of users, the sub-bureau prediction predicts whether users will repost a specific weibo, The focus of sub-bureau prediction is to predict the individual repost behavior of users by analyzing their behavioral characteristics, such as their history list of weibo repost and the preferences of repost behavior. Global prediction refers to the prediction of the repost scale and propagation scope of a specific weibo, as well as its popularity on the platform of social networks. The global repost prediction is mainly related to the influence of users who has published the weibo, the degree of activity of their fans, content the release time of the published weibo.

For the weibo information in the displayed weibo list, users mainly have three behaviors: repost, ignoring and unreceiving. Repost behavior refers to the behavior that a user clicks the repost button to repost his interested or favorite weibo when reading the weibo information from the list. Ignoring behavior refers to the behavior that the user does not take any action on a certain weibo after reading it. The unreceived behavior refers to the behavior that the user is not online when the new weibo list appears, and the mentioned weibo list has been updated when the user goes online again, but the user is completely unaware of the weibo list hat have already appeared.

* Corresponding Author

The relations of repost in social networks can be represented by directed graphs [5] G=(U, E), where, U represents users in social networks, and E represents the repost relations between users. The outgoing degree of U is the number of the followers of U, the incoming degree of U is the number of the fans of U. From the perspective of graph theory, degree of nodes, average distance and clustering coefficient of G are all important indicators for analyzing social networks. The degree value of a node is related to its influence in the social network. The greater the degree value of node, the greater its influence and the stronger its ability to disseminate information. Average distance refers to the shortest path length between two nodes and describes the basic characteristics of the network. The clustering coefficient can reflect the degree of node aggregation in the network. The network has the relative complexity of time and space. With the change of time, the space of the network also changes constantly, showing the complexity of mutual influence and interaction between nodes. Therefore, the research on network structure is helpful to improve the accuracy of predicting weibo repost.

As one of the branches of mathematical statistics, time series analysis method [3] reveals the dynamic structure and rules of the system based on mathematical statistics and probability theory. Time series analysis [6-8] is a statistical method to reveal the dynamic structure and rules of a system based on dynamic data. The time series analysis in the research of weibo repost is to analyze the time characteristics of weibo repost amount and master the repost law of weibo. After reasonable data changes, the time series of weibo repost behaviors can be regarded as the superposition of trend items, period items and random noise parts. The main work of time series analysis is to analyze these three parts of time series and establish a mathematical model that can accurately reflect the system structure and data dependency. So as to further predict the repost behavior of weibo. Nowadays, Autoregressive Integrated Moving Average Model (ARMIA) has been proved an useful model to analyze the time series. However, data analysis shows that the time series analysis results of weibo repost by using the ARMIA model will generate noises, these noises may come from the false repost of the original weibo or multiple reposts and so on. In the actual analysis, the noises should be removed. In terms of signal processing, the wavelet transform can remove the noises of signal. Data analysis has shown that the characteristics time series repost of weibo are similar to the signal propagation, they both have ‘burrs’ on the time series curve, and these burrs are the noises which will influence the accuracy of predicting weibo repost by time series analysis.

From what have been discussed above, in this paper, we try to propose an improved method by introducing the wavelet transform mechanism based on the ARMIA model [9]. In particular, the method of zeroing the high frequency coefficients of multi-scale analysis in wavelet transform is adopted to reduce the noise of the forwarding amount. We propose an improved data processing method based on wavelet theory to remove noise in time series and to enhance the ability of predicting the weibo reposts.

2 Model

2.1 ARIMA Model

Basic idea of Autoregressive Moving Average Model (ARMA) [10-11] is: some of the time series is a set of random variables depends on the time, although a single sequence value of the time sequence is uncertain, changes in the sequence has a certain regularity. The corresponding mathematical model can be used to describe the regularity in an approximate way. The final goal of ARIMA is to reach the minimum variance under the optimal prediction by analyzing the structure and characteristics of time series.

ARIMA [12-13] is consisted of AR model and MA model. AR model is a p order autoregressive process by using itself as the regression variables, then the time series Y in the time point t can be expressed as:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \quad \textbf{(1)}$$

We can get the conclusion that time series value of Y_t is the linear combination of its pre-p order time series (e.g. Y_t... Y_{t-p}) and white noise (e_t). $\phi_1, \phi_2, \cdots, \phi_p$ are the regression coefficients, and white noise obey the normal distribution with mean value 0 and variance value σ^2. If we define $B^k Y_t = Y_{t-k}$, then formula (1) will be updated as:

$$Y_t = \phi_1 B^1 Y_t + \phi_2 B^2 Y_t + \cdots + \phi_p B^p Y_t + e_t \tag{2}$$

Suppose $\varphi(B) = 1 - \varphi_1 B^1 - \varphi_2 B^2 - \cdots - \varphi_p B^p$, then the formula (2) can be abbreviated as

$$\varphi(B) Y_t = e_t \tag{3}$$

MA model assumes that the time series Y_t can be linearly defined by the random error, it can be expressed as the following formula:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{4}$$

where e_t is the white noise sequence and $\theta_1, \theta_2, \cdots, \theta_q$ are the moving average coefficients (parameters). By introducting B^k, and defining $\theta(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - \cdots - \theta_q B^q$, formula (4) can be abbreviated as

$$Y_t = \theta(B) e_t \tag{5}$$

Then AR model and MA model can be combined as ARMA model:

$$Y_t = \phi_1 B^1 Y_t + \phi_2 B^2 Y_t + \cdots + \phi_p B^p Y_t + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{6}$$

where p and q represent the value of order of AR and MA separately. Future, the formula of ARMA will be abbreviated as

$$\varphi(B) X_t = \theta(B) e_t \tag{7}$$

Assuming that the ARMA (p, q) model is stable, if the time series $\{Y_t\}$ reach to a stable ARMA(p, q) sequence through $W_t = \nabla^d Y_t$, then time series $\{Y_t\}$ is Autoregressive Integrated Moving Average Model (ARIMA), where Δ^d the differential factor, d is the differential value. ARIMA can be expressed as ARIMA(p, d, q). When d=1, $W_t = Y_t - Y_{t-1}$, the ARIMA can be expressed as

$$W_t = \phi_1 W_{t-1} + \phi_2 W_{t-2} + \cdots + \phi_p W_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \tag{8}$$

Time series can be expresses as:

$$\begin{aligned} Y_t - Y_{t-1} = {} & \phi_1 (Y_{t-1} - Y_{t-2}) + \phi_2 (Y_{t-2} - Y_{t-3}) + \cdots + \phi_p (Y_{t-p} - Y_{t-p-1}) \\ & + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \end{aligned} \tag{9}$$

For more details, please see the reference [14].

2.2 Time Series Prediction Based on ARIMA

Establish a complete ARIMA model and predict the time series sequence including the following seven steps.

(1) Adjust the original time series to generate the appropriate time series: We use data aggregate method to count data according to time unit, such as counting the number of weibo reposts per hour (or 30 minutes). In this step, time unit selection is the key point. On one hand, the data of time unit should reflect the change of the trend efficiently, on the other hand, keeping enough data of time unit will improve the efficiency of data prediction.

(2) Inspect the stationarity of time series: There are many tools to achieve this goal, such as scatter diagram, ADF unit root test [15]. If the time series is not stable, then we use poor differentiation method to handle the data and to convert the initial time series data to a stationary time series sequence.

(3) Preliminary recognition of the model: Computing the autocorrelation function (AFC) and partial autocorrelation function (PAFC) [16] using the following formuas:

$$\text{AFC: } r_k = \frac{\sum_{t=k+1}^{n}(\mathrm{Y}_t - \overline{\mathrm{Y}})(\mathrm{Y}_{t-k} - \overline{\mathrm{Y}})}{\sum_{t=1}^{n}(\mathrm{Y}_t - \overline{\mathrm{Y}})^2}, \quad k = 1, 2\cdots \tag{10}$$

$$\text{PAFC: } \phi_{kk} = Corr(\mathrm{Y}_t, \mathrm{Y}_{k-t} \mid \mathrm{Y}_{t-1}, \mathrm{Y}_{t-2}, \cdots, \mathrm{Y}_{t-k+1}) \tag{11}$$

Then we will choose the suitable order p and q (see formula (2) and (4)), based on the following table.

Table 1. The reference of model order

	AR(p)	MA(q)	ARMA(p, q)
ACF	long-tail point	long-tail point after q order	long-tail point
PACF	long-tail point after p order	long-tail point	long-tail point

(4) Check the order(p, q, d) of the model: We use Akaika information criterion(AIC), Schwarz Bayes Criterion(SBC), Hannan-Quinn Information Criterion(HQIC) to check the order of the model [17-20].

AIC is defined as

$$\mathrm{AIC} = 2k - 2\ln(\mathrm{L}) \tag{13}$$

where L is the maximum likelihood function of the model, k is the number of parameters which need to be estimated. The smaller the value of AIC, the better selection of the model's order.

SBC is defined as

$$SBC = k\ln(\mathrm{n}) - 2\ln(\mathrm{L}) \tag{14}$$

where n is the number of observations. When the number of observations n is large enough, the order evaluated by AIC is usually higher than that of real model, and the best model determined by SBC is closer to real data. Similary, the smaller the value of SBC, the better selection of the model's order.

HQIC is defined as

$$\mathrm{HQIC} = 2k\ln(\mathrm{n}) - 2\ln(\mathrm{L}) \tag{15}$$

Similary, the smaller the value of HQIC, the better selection of the model's order.

(5) Model parameters estimation: We use the maximum likelihood estimation to estimate model parameters, because the maximum likelihood estimation method has the advantage of high accuracy. Some other estimation methods, such as torque estimation and the least squares estimation, also cen be utilized.

(6) Model parameters estimation: Test the goodness of the fitting of model through residual analysis. Repeat the above steps until the best fitting model is determined.

(7) Prediction: Predict the time series sequence in the future period of time by using the best model, which is achieved from the mentioned six steps. We use the minimum mean square error of prediction in this paper.

For more intuitive understanding of the whole process, we use the following flow chart to show how the above seven steps work.

3 Real Data Analysis

3.1 Dataset

In this paper, we will use the data of Sina weibo to analyze the weibo reposts. The data used in this section are obtained through the API interface provided by Sina weibo. In order to study the behavior of weibo reposts and predict the scale of weibo reposts, we chose the representative weibo event which has great influence and has been reposted more than 200,000 reposts number to illustrate the phenomenon. weibo news: Super star Zhou announced his marriage through weibo on November 10, 2015 at 6:54 p.m. For this weibo, we counted its specific repost number in 14 days, the statistical results is shown in Fig. 2.

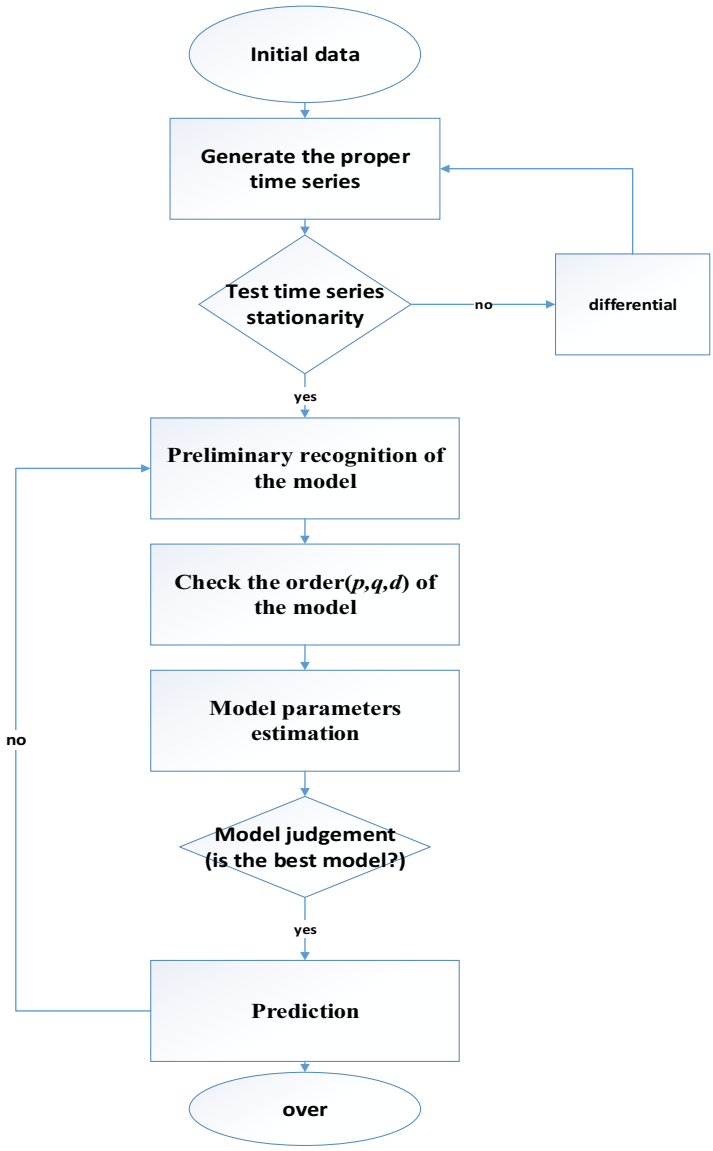


Fig. 1. Flow chart of prediction based on ARIMA

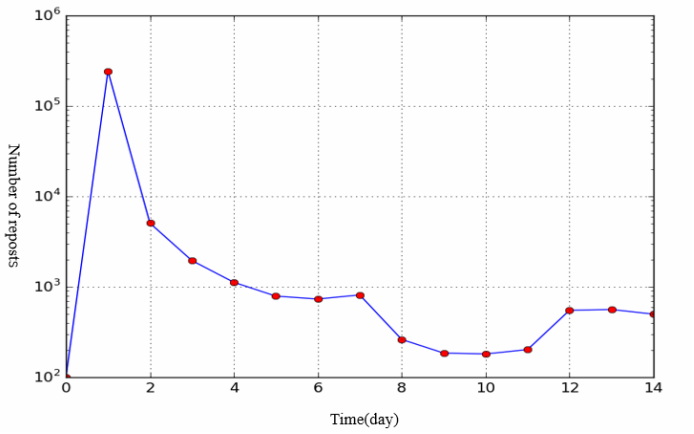
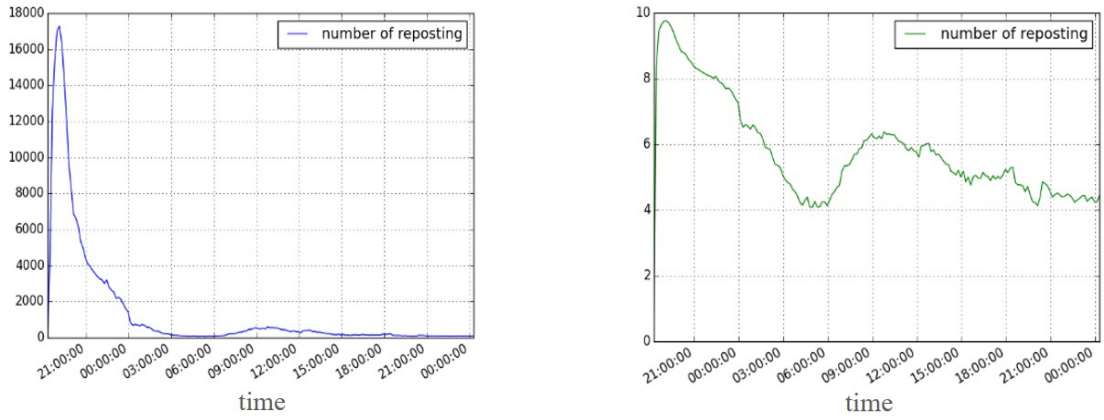


Fig. 2. Statistic of weibo reposts of the news "Super star Zhou has married with his wife" in 14 days

We can see that repost behiever mainly happens within two days after the launch of this weibo, or even within one day. The subsequent reposts number is maintained within the magnitude of 10^2-10^3. This phenomenon shows that weibo content are of strong feature of news timeliness.

More specifically, further analysis of weibo reposts data shows that weibo reposts are mainly concentrated in 30 hours after the release of weibo. Then we will focus on the analysis of weibo reposts data within the mentioned 30 hours.

Fig. 3 shows that the data in 10 minutes is a time node to count the amount of theweibo reposts. There are a total of 180 valid data in the figure. It can be seen from the Fig. 3(a) that the reposting amount per unit time reaches the peak at the moment after the weibo is released. According to the analysis, the number of users' followers plays a crucial role in the reposting of weibo, because only fans can see the latest microblog of the blogger in the first time. The ordinate of Fig. 3(b) is the logarithmic coordinate with base e as the base. It can be seen from the figure that after the logarithm of the data is taken, the change of the data presents a relatively stable trend. The coordinate transformation in this step can be regarded as data preprocessing. Then we use the mentioned ARIMA model (including seven steps) to further analyse data.



(a) (b)

Fig. 3. Distribution of weibo reposts data within 30 hours

First, we will take the data after logarithm for a difference processing, and the processing results are shown in Fig. 4.

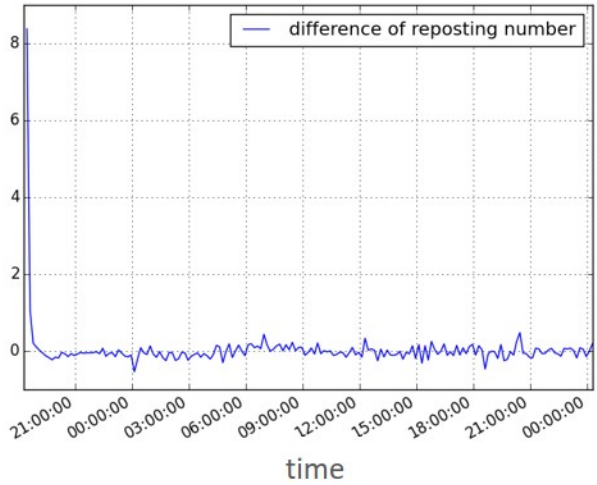


Fig. 4. The processing result of time series difference

It can be seen from the figure, after the first difference, the whole time series has appeared very stable, except for the time series of the beginning time. That is because the reposts reached a considerable number within the instant of weibo's release. The result of primary difference shows that the original series has good stability. Then we will compute ACF and PACF. The results are shown in Fig. 5(a) (ACF) and Fig. 5(b) (PACF). As can be seen from the figure, with the increase of lag length K, AFC decayed

slowly linearly, while PACF was significantly zero after k=5. According to Table 1, AR(5) model should be chosen as the analysis model.

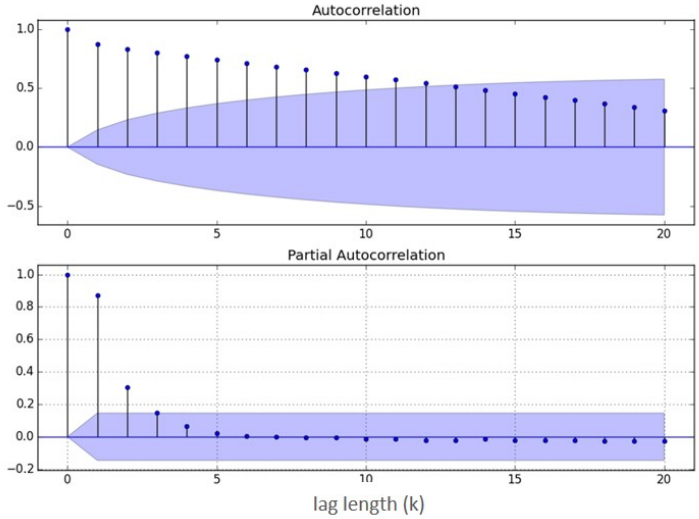


Fig. 5. The results of ACF and PACF

However, in the actual calculation process, on the one hand, it is hoped that the lag length can large enough to reflect the dynamic characteristics of the construction model; But on the other hand, the longer the lag length, the more parameters should be estimated in the model, and the more degrees of freedom will be lost. So we need to find an equilibrium between lag and freedom. This section combines several criteria to determine the order of the model. On the basis of ACF and PACF, we expand the selection range of model order to find a more suitable model. We successively calculated AIC, SBC and HQIC of different models. Each final result is averaged over ten calculations, and the results were arranged in the following table. As can be roughly seen from Fig. 6, with the increase of the model order, the value of AIC, SBC and HQIC are increasing, which indicates that the problem of over-fitting becomes more serious with the increase of the model order.

ARMA	AIC	SBC	HQIC	ARMA	AIC	SBC	HQIC
ARMA30	325.7652	341.7577	332.2489	ARMA80	334.1604	366.1454	347.1278
ARMA31	327.3475	346.5385	335.128	ARMA81	342.57	377.7534	356.8341
ARMA32	327.4221	349.8116	336.4993	ARMA82	342.7913	381.1733	358.3522
ARMA40	327.6465	346.8374	335.4269	ARMA83	335.2246	376.805	352.0822
ARMA41	326.7156	349.1051	335.7928	ARMA84	344.015	388.7939	362.1693
ARMA42	330.1599	355.7479	340.5338	ARMA85	346.8817	394.8592	366.3328
ARMA43	328.7298	357.5163	340.4004	ARMA86	340.7929	391.9689	361.5407
ARMA50	329.5531	351.9426	338.6303	ARMA87	341.983	396.3574	364.0275
ARMA51	337.0619	362.6499	347.4358	ARMA90	336.1373	371.3207	350.4014
ARMA52	328.0307	356.8172	339.7014	ARMA91	338.1447	376.5267	353.7056
ARMA53	328.2928	360.2778	341.2602	ARMA92	339.3035	380.884	356.1611
ARMA54	332.3294	367.5128	346.5935	ARMA93	377.1033	421.8823	395.2577
ARMA60	331.5429	357.1308	341.9168	ARMA94	365.1659	413.1434	384.617
ARMA61	332.409	361.1955	344.0797	ARMA95	386.5708	437.7468	407.3187
ARMA62	331.51	363.4949	344.4774	ARMA97	336.8569	394.4299	360.1982
ARMA63	333.5501	368.7336	347.8142	ARMA98	345.0521	405.8235	369.6901
ARMA64	337.1901	375.5721	352.751	ARMA100	337.9424	376.3243	353.5032
ARMA65	336.7883	378.3688	353.646	ARMA101	339.6896	381.2701	356.5472
ARMA70	332.3336	361.1201	344.0042	ARMA102	341.6697	386.4486	359.824
ARMA71	334.1481	366.1331	347.1155	ARMA103	339.8546	387.8321	359.3057
ARMA72	332.0474	367.2309	346.3115	ARMA104	347.466	398.6419	368.2138
ARMA73	338.5988	376.9808	354.1597	ARMA105	342.4693	396.8438	364.5139
ARMA74	338.5154	380.0959	355.373	ARMA106	341.099	398.672	364.4403
ARMA75	335.2412	380.0202	353.3956	ARMA107	342.5023	403.2737	367.1403
ARMA76	463.9269	511.9044	483.378	ARMA108	353.9657	417.9356	379.9004

Fig. 6. Results of AIC, SBC and HQIC

According to AIC, SBC and HQIC values, we respectively found the relative top 18 models and made statistics in the below Fig. 7. It can be seen from Fig. 7 that although the ranking of models is different,

the overlap degree is large. ARMA (4, 3) (5, 2) (5, 3) (6, 0) was selected as the order number of the model after comprehensive consideration according to Fig. 7.

rank	AIC	SBC	HQIC	rank	AIC	SBC	HQIC
1	ARMA30	ARMA30	ARMA30	10	ARMA42	ARMA43	ARMA53
2	ARMA41	ARMA31	ARMA31	11	ARMA62	ARMA53	ARMA60
3	ARMA31	ARMA40	ARMA40	12	ARMA60	ARMA70	ARMA70
4	ARMA32	ARMA41	ARMA41	13	ARMA72	ARMA61	ARMA61
5	ARMA40	ARMA32	ARMA32	14	ARMA54	ARMA51	ARMA62
6	ARMA52	ARMA50	ARMA50	15	ARMA70	ARMA62	ARMA72
7	ARMA53	ARMA42	ARMA52	16	ARMA61	ARMA71	ARMA54
8	ARMA43	ARMA52	ARMA43	17	ARMA63	ARMA80	ARMA71
9	ARMA50	ARMA60	ARMA42	18	ARMA71	ARMA72	ARMA80

Fig. 7. Rank of different modeld according to the results of AIC, SBC and HQIC

To further select the parameters of the model. We performed residual analysis on the selected model. The specific formula is expressed as

$$\hat{e}_t = Y_t - \hat{\pi}_1 Y_{t-1} - \hat{\pi}_2 Y_{t-2} - \hat{\pi}_3 Y_{t-3} - \cdots \tag{16}$$

where π is not directly estimated, but implicitly calculated as a function of ϕ and θ. As you can see from Fig. 8, the difference of four residual figures is not very big, the main difference is concentrated in the beginning of the time series. In the actual reposting process, the user (superstar) has many fans and followers, expecially including a large number of media users, so when the user once released an explosive news, the news of instantaneous forward quantity can reach a peak. The occurrence of this phenomenon basically has no time transition (happened in a very short time), which makes it difficult to fit the model and makes the residual analysis appear unreasonable at the beginning.

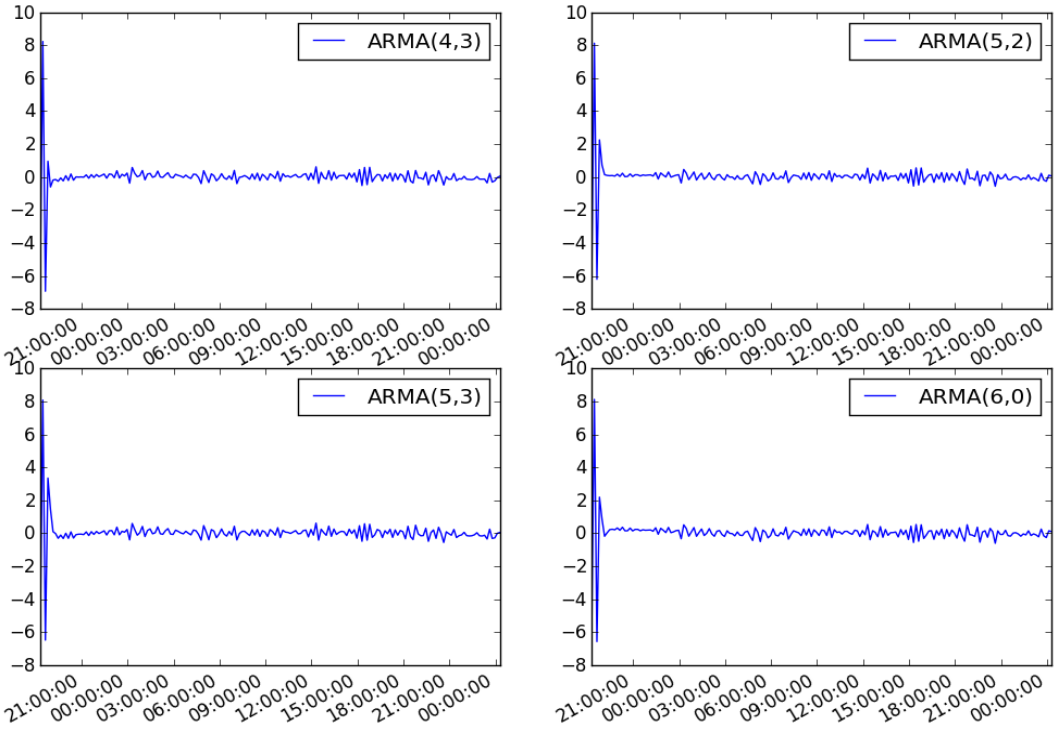


Fig. 8. Comparison of residual results of ARMA(4, 3) ARMA(5, 2) ARMA(5, 3) ARMA(6, 0)

To further analyze the properties of residuals, Fig. 9 shows the quantile-quantile plot of the four models. It can be seen from the figure that although the general distribution is the same, the residual analysis of ARMA(4, 3) performs better.

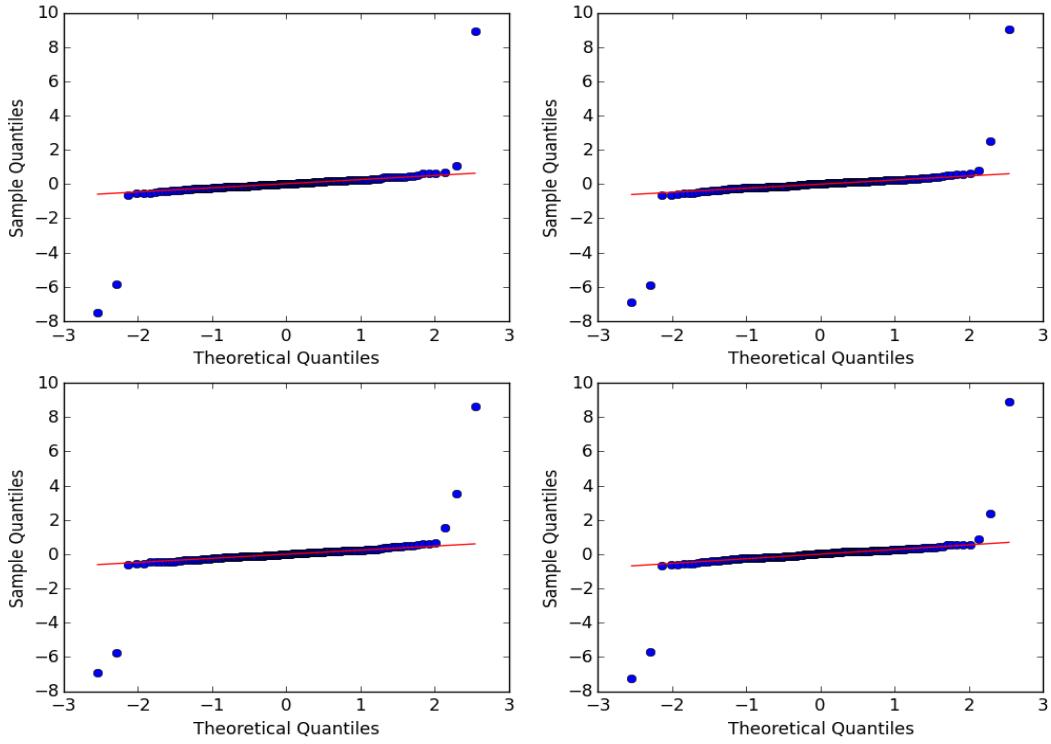


Fig. 9. Comparison of quantile - quantile of ARMA(4, 3) ARMA(5, 2) ARMA(5, 3) ARMA(6, 0)

Summarize the process: we draw a model order for AR (5) by computing ACF and PACF. In order to find more suitable model parameters and to avoid the error caused by a single theory, then we use AR (5) model as the middle point and adopt the method of trial and error one by one to each side extension, and meanwhile to calculate the values of AIC, SBC and HQIC. Then we select some model orders with high fitting degree. In order to further check the fitting degree of the model, the quantile-quantile graph was used to analyze the residual of the model, so as to determine the basic order of the model.

3.2 Prediction of Weibo Repost Based on ARIMA

According to the model selected above, we began to predict the reposting time series of weibo. Firstly, 180 time sieres data were divided into training group and test group. The first 100 time series data is selected as the training group to predict the following 80 time series data. Mean Squared Error (MSE) will be used as the evaluation indicator. Each experiment was repeated 20 times independently, and the average value was taken as the result of model prediction.

Fig. 10 and Fig. 11 show the difference between the average prediction results and the test group data. The blue line represents the actual reposting number, and the green line represents the predicted reposting number. It can be seen from the figure that the predicted results always fluctuate around the real data, and the predicted results are basically consistent with the actual results, especially at the turning point of the time series. Intuitively, it is felt that there is little difference about the prediction results of the four models, then we will calculate MSE between predicted values and real values. It can be seen from the results of Table 2 that the prediction ability of the model ARMA(4, 3) is better. Although there is little difference in MSE value, considering that the calculated data are obtained by logarithm of the original data, even if there is little difference in MSE, there will be great difference in the actual prediction effect.

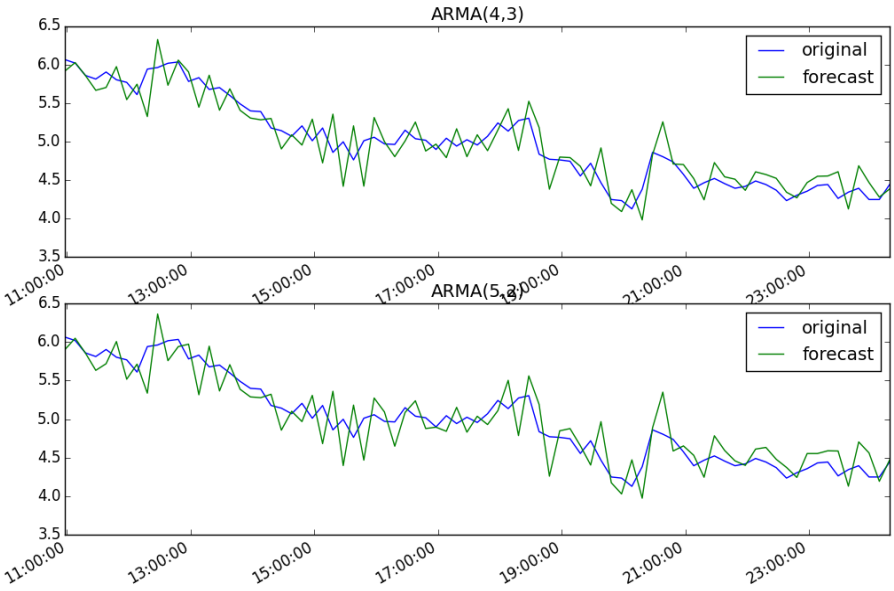


Fig. 10. Prediction results of ARMA(4, 3) ARMA(5, 2)

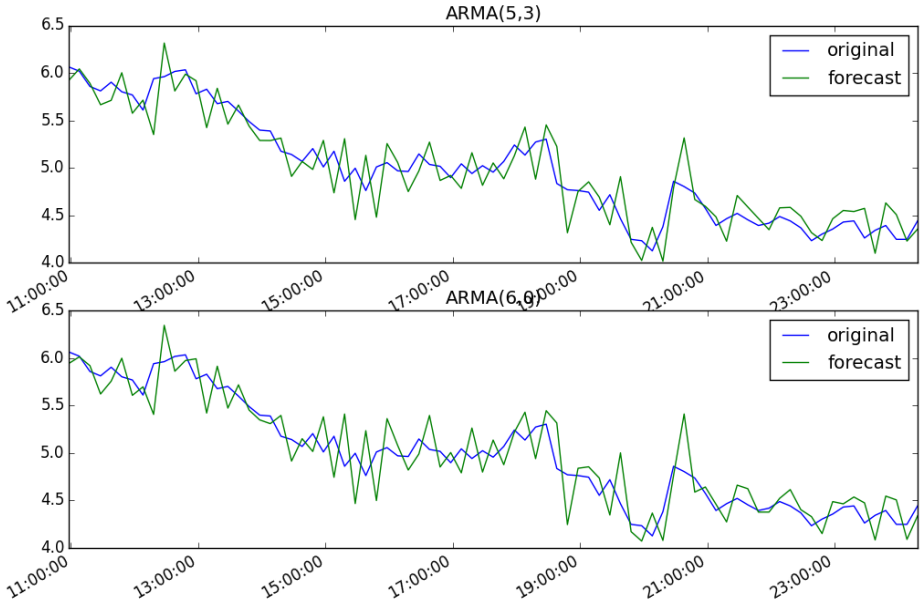


Fig. 11. Prediction results of ARMA(5, 3) ARMA(6, 0)

Table 2. MSE of ARMA(4, 3) ARMA(5, 2) ARMA(5, 3) ARMA(6, 0)

model	MSE
ARMA (4, 3)	0.062738839725965306
ARMA (5, 2)	0.062935156363783079
ARMA (5, 3)	0.070400963035574748
ARMA (6, 0)	0.065975147982828247

4 Model Improvement

4.1 Wavelet Analysis Theory

We will use wavelet analysis theory [21-22] to improve the model. We set $\forall \psi \in L^2(R)$, $\hat{\psi}(\omega)$ is the fourier transform of $\psi(t)$, if $\hat{\psi}(\omega)$ satisfies the constraints

$$c_{\psi} = \int_{-\infty}^{+\infty} \frac{|\hat{\psi}(\omega)|^2}{|\omega|} d\omega < +\infty \tag{17}$$

$\psi(t)$ is called a wavelet generating function or a basic wavelet. The wavelet generating function can be shifted and scaled to obtain a series of sub-wavelets:

$$\psi_{a,b}(t) = |a|^{-\frac{1}{2}} \psi(\frac{t-b}{a}) \quad a \neq 0, b \in R \tag{18}$$

where *a* is the scale parameter, the subwavelet can stretch out and draw back according to *a, b* is the shift parameter, the subwavelet can shift according to *b.* Wavelet analysis is a transformation of time domain and frequency domain, the scale parameter determines the spectrum transformation of $\psi(\frac{t-b}{a})$, which is equivalent to filtering the signal with the same shape, different bandwidth and main frequency. When a>1, the wavelet $\psi(t)$ will stretch out, When a < 1, the wavelet $\psi(t)$ will draw back. But the effect on $\hat{\psi}(\omega)$ is just the opposite. Therefore, the change of *a* will affect the temporal resolution and frequency resolution of the signal. When b > 0, the wavelet $\psi(t)$ will move right, when b < 0, the wavelet $\psi(t)$ will move left. Therefore, by changing *b*, the region where the signal can be analyzed will be selected.

For any signal $f(t)$ is a space signal with finite energy, its continuous wavelet transform is defined as:

$$W_f(a,b) = < f, \psi_{a,b} > = |a|^{-\frac{1}{2}} \int_{-\infty}^{+\infty} f(t) \overline{\psi}(\frac{t-b}{a}) dt \tag{19}$$

Discrete wavelet transform refers to the discretization of scale parameter *a* and shift parameter *b* rather than the time variable *t*. We set $a = a_0^m$, $b = nb_0a_0^m$, $a_0 > 1$, $b_0 \in R$, then discrete wavelet transform of $f(t)$ is:

$$W_f(m,n) = < f, \psi_{m,n} > = a^{-\frac{m}{2}} \int_{-\infty}^{+\infty} f(t) \overline{\psi}(a_0^{-m}t - nb_0) dt \tag{20}$$

However, signals or information sources in real problems are often accompanied by varying degrees of noise. In a broad sense, all the components except the active components are collectively referred to as noise. A signal or information source with noise can be expressed as:

$$s(t) = f(t) + n(t) \tag{21}$$

Where $s(t)$ is the source signal, $f(t)$ is the effective component of the signal, and $n(t)$ is the noise part. Only the noise is removed, the most useful and core part of a signal or information can be extracted to improve the efficiency of information processing. There are many effective methods of noise reduction by wavelet transform. We use the high frequency zero noise reduction method of wavelet to carry out multi-scale analysis of information [23]. Below is a schematic diagram of multi-scale analysis of signals.

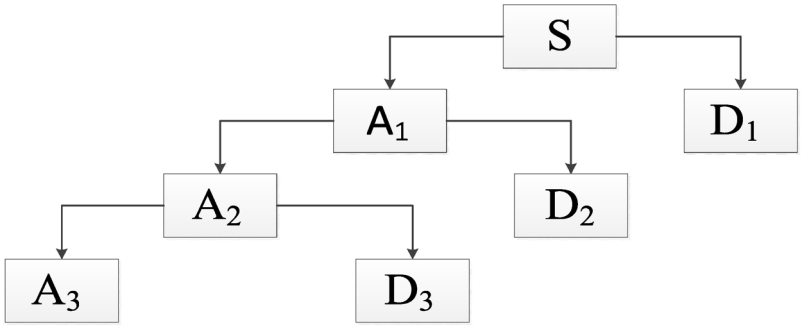


Fig. 12. Three - layer multi - scale analysis of signal

Where S is the signal to be analyzed, A is the approximate part, the large scale low frequency part of the corresponding signal, D is the detail part, and the small scale high frequency part of the corresponding signal.

4.2 Model Improvement Based on Multi-scale Analysis

In Fig. 3, the statistical unit of reposting amount is 10 minutes, that is, the reposting amount within 10 minutes is evenly distributed. However, in the actual situation, the reposting amount varies from minute to minute. The blue curve in Fig. 13 is the reposting volume distribution diagram using 5 minutes as the statistical unit. It can be seen from the figure that the curve has a lot of burrs and becomes very unsmooth, if the statistical unit is further reduced, the burr phenomenon will be more serious. From the perspective of signal processing, these burrs are noise, which will affect the prediction of overall trend of the future curve. The noise in the reposting quantity may come from the false reposting or the reposting of the original weibo for many times and so on. In order to improve the accuracy of repost prediction, these noises should be removed. Combined with the characteristics of the method of wavelet noise reduction in the previous section, we will adopt the high frequency zero noise reduction method of multi-scale analysis to reduce the noise of the repost amount. In high frequency of zero noise reduction, effective part mainly concentrated in the low frequency part of signal, only a few effective part in the high frequency part, and then a lot of noise is concentrated in the high frequency part, so the signal through the low-pass and high-pass filter, the signal is divided into low frequency part and high frequency part, the high frequency coefficients are zero, and the low frequency part break down again, so, reach the multi-scale analysis of signals, finally according to the low frequency coefficients of signal reconstruction.

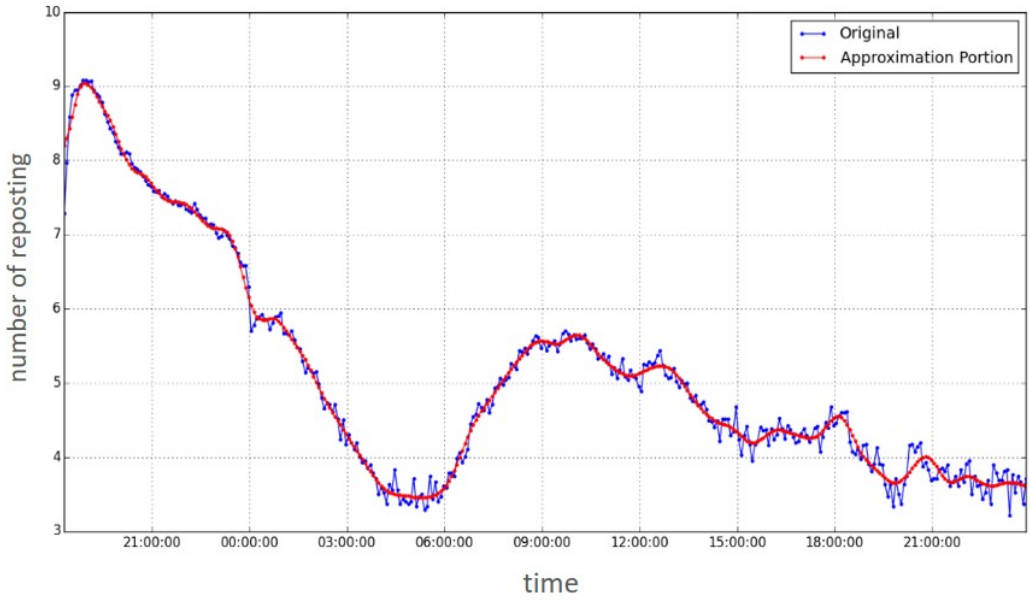


Fig. 13. The statistic of weibo reposts in 5 minutes

According to different application characteristics of the wavelet application in multi-scale analysis and noise reduction, Daubeehies (dbN) wavelet is the best choice, and db8 is selected as the use of wavelet function. Considering the comprehensive performance, the repost amount of weibo is finally analyzed at three layers multiple scales. The high frequency coefficient is set to zero, and then the forward amount is reconstructed according to the low frequency coefficient A_3 to achieve the purpose of noise removal.

The red curve in Fig. 13 is the distribution diagram of weibo reposting after reducing noise. It can be seen from the figure that the low-frequency part of the weibo reposting maintains the trend and characteristics of the original data, and it covers the most effective components of the data, and will be helpful to provide a clear weibo reposting trend chart to researchers. Fig. 14, Fig. 15 and Fig. 16 show the high frequency part of the data with high oscillation frequency and small oscillation amplitude. It contains a large number of unsteady state parts, namely, the noise part. After noise reduction, the data is

smooth, and the burrs in the data are eliminated, so that the curve has an obvious trend of change, which lays a good foundation for improving the prediction accuracy.

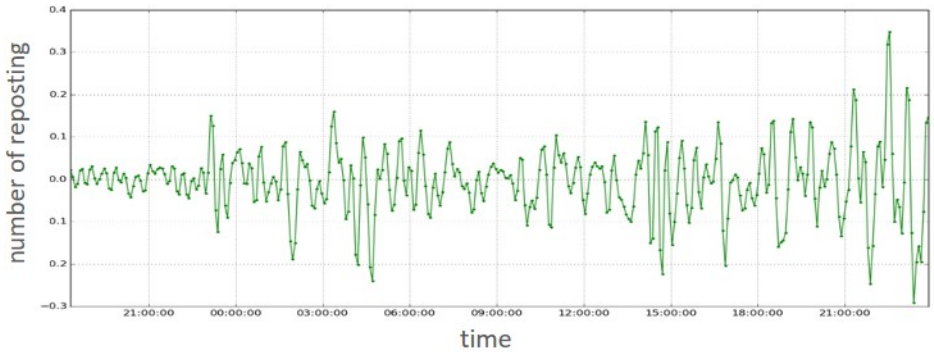


Fig. 14. The high frequency part of the number of reposts D1

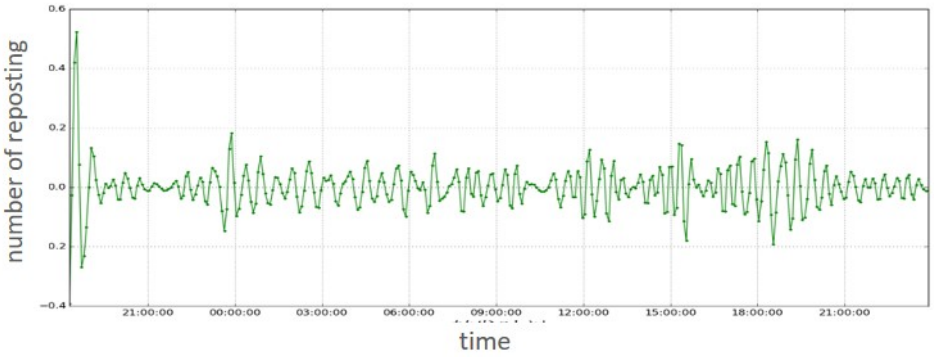


Fig. 15. The high frequency part of the number of reposts D2

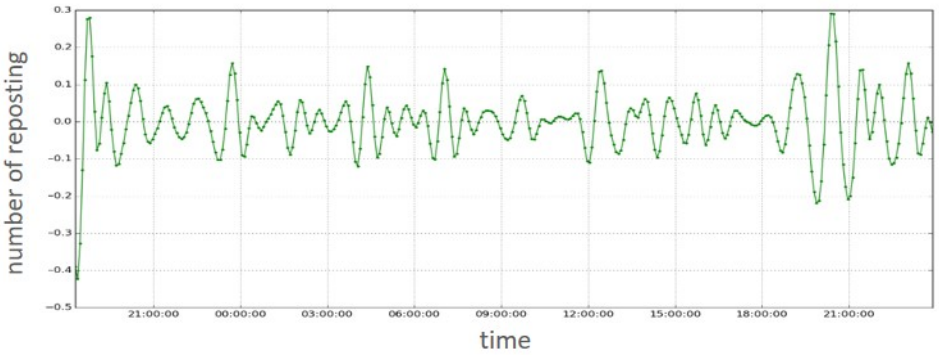


Fig. 16. The high frequency part of the number of reposts D3

Model prediction in the last section makes curve smoothing by changing the statistics unit, in order to achieve the role of primary noise reduction. This method makes the reposting amount of statistical unit as fixed value and cannot extract the effective part of it. Meanwhile, it will reduce the accuracy of prediction. Through the method of wavelet noise reduction can both keep effective part of the data, and can remove the noise in the data section. It is a kind of give attention to both efficiency and quality of the good measures.

5 Conclusion and Future Work

In Sina weibo, celebrity effect is very obvious, and celebrities' micro-blogs usually attract huge attention and reposts. We use super star Zhou's weibo content to analyze the time series data of weibo, ARIMA

model was utilized to predict Zhou's weibo reposts in the future time. In order to improve the prediction effect of the model, wavelet analysis method is used to remove the noise part of the data in the pre-processing stage. Experimental results show that multi-scale analysis combined with high-frequency zeroing can retain the effective part of the data and remove the noise part of the high-frequency, making the curve smooth. When the statistical unit is continuously reduced, the wavelet denoising method can not only retain the details of the data, but also remove the corresponding noise part. Only the data that removes noise can improve the performance of weibo reposts prediction. The selection of wavelet and the selection of noise reduction methods will be the future questions for the time series data processing and analysis of weibo.

Acknowledgements

This work has been supported by the National Key Research and Development Program of China (grant number 2018YFC0831703), Fundamental Research Funds for the Central Universities (grant number: 2020JBM002), and China Postdoctoral Science Foundation (grant number: 2018M641172).

References

[1] H.L. Zhu, M. Wang, A novel reposting prediction method based on quantified microblog hotness in sina weibo, in: Proc. International Conference on Computer Science and Application Engineering (Csae), 2017.

[2] G.X. Luo, Y. Liu, Z.Y. Zhang, A novel model for weibo reposts prediction by using generic based segmented BPNN, Journal of Internet Technology 19(2018) 1293-1301

[3] K. Zhao, Y.Q. Zhang, B.G. Li, C.F. Zhou, Repost number prediction of micro-blog on sina weibo using time series fitting and regression analysis, in: Proc. 2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI), 2015.

[4] W.Y. Yao, P.F. Jiao, W.J. Wang, Y.H. Sun, Understanding human reposting patterns on Sina Weibo from a global perspective, Physica a-Statistical Mechanics and Its Applications 518(2019) 374-383.

[5] G.Z. Zhang, Y.C. Sun, M.L. Xu, R.F. Bie, Weibo clustering: a new approach utilizing users' reposting data in social networking services, Computer Science and Information Systems 11(2014) 1157-1172.

[6] K. Lyudmyla, B. Vitalii, R. Tamara, Fractal Time Series Analysis of Social Network Activities, in: Proc. 2017 4th International Scientific-Practical Conference Problems of Infocommunications-Science and Technology (Pic S&T), 2017.

[7] K. Nusratullah, S.A. Khan, A. Shah, W.H. Butt, Detecting changes in context using time series analysis of social network, in: Proc. 2015 Sai Intelligent Systems Conference (Intellisys), 2015.

[8] J.A. Danowski, N. Cepela, Automatic mapping of social networks of actors from text corpora: time series analysis, 2009 International Conference on Advances in Social Networks Analysis and Mining, 2009.

[9] N.M. Ariff, N.H. Zamhawari, M.A. Abu Bakar, Time series ARIMA models for daily price of palm oil, in: Proc. 2nd ISM International Statistical Conference, 2014.

[10] A.Y. Zheng, W.M. Liu, F.G. Zhao, Double trends time series forecasting using a combined ARIMA and GMDH model, in: Proc. 2010 Chinese Control and Decision Conference, 2010.

[11] S.S. Pappas, L. Ekonomou, D.C. Karamousantas, G.E. Chatzarakis, S.K. Katsikas, P. Liatsis, Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models, Energy 33(2008) 1353-1360.

[12] K. Kumar, V.K. Jain, Autoregressive integrated moving averages (ARIMA) modelling of a traffic noise time series, Applied Acoustics 58(1999) 283-294.

[13] M.H. Amini, A. Kargarian, O. Karabasoglu, ARIMA-based decoupled time series forecasting of electric vehicle charging demand for stochastic power system operation, Electric Power Systems Research 140(2016) 378-390.

[14] X.J. Ren, X. Chen, Discovery and dynamic prediction of user's interest based on ARIMA, in: Proc. 2017 Portland International Conference on Management of Engineering and Technology, 2017.

[15] J.H. Lopez, The power of the ADF test, Economics Letters 57(1997) 5-10.

[16] A. Inoue, AR and MA representation of partial autocorrelation functions, with applications, Probability Theory and Related Fields 140(2008) 523-551.

[17] S.E. Lazic, Model based inference in the life sciences: a primer on evidence, Journal of the Royal Statistical Society Series a-Statistics in Society 174(2011) 506-506.

[18] E.J. Wagenmakers, Model selection and multimodel inference: a practical information-theoretic approach., Journal of Mathematical Psychology 47(2003) 580-586.

[19] M.R.E. Symonds, A. Moussalli, A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion, Behavioral Ecology and Sociobiology 65(2011) 13-21.

[20] M.J. Whittingham, P.A. Stephens, R.B. Bradbury, R.P. Freckleton, Why do we still use stepwise modelling in ecology and behaviour?, Journal of Animal Ecology 75(2006) 1182-1189.

[21] A.A. Davydov, Wavelet analysis of social processes, Sotsiologicheskie Issledovaniya (2003) 89-101.

[22] G. Wittemyer, L. Polansky, I. Douglas-Hamilton, W.M. Getz, Disentangling the effects of forage, social rank, and risk on movement autocorrelation of elephants using Fourier and wavelet analyses, in: Proc. National Academy of Sciences of the United States of America, 2008.

[23] Z. Liu, J. Tang, M.Y. Jia, X.J. Zhou, Analysis method of multi-scale frequency spectral latent feature based on the mutual information, in: Proc. 2015 3rd International Conference on Machinery, Materials and Information Technology Applications, 2015.