ISSN: 2664-9640

Research on traffic flow prediction Technology based on big data

Jingchen Liu

School of Transportation, Shanghai Maritime University, Shanghai 200000, China.

Abstract

With the development of social economy, urban traffic congestion becomes more and more serious, which brings great inconvenience to residents' daily life. In order to alleviate traffic congestion, scholars have conducted in-depth studies and applied various models to short-term traffic flow prediction, such as time series model, non-parametric regression model and neural network model.At present, in the short-term traffic flow prediction, there are still few models using big data technology to predict traffic flow. This paper focuses on the exponential smoothing method, analyzes the advantages and disadvantages of the first exponential smoothing and the third exponential smoothing, and draws out the shortcomings of the third exponential smoothing method, so as to put forward the improvement scheme, and combine it with the big data technology to verify the scheme and prove its superiority.

Keywords

Big data, MapReduce, exponential smoothing, dynamic model.

1. Introduction

The International Data Corporation released a global Data total of 10ZB in 2016 and predicted it will reach 35ZB in 2020[1]. In this data explosion environment, there are many sources of massive data, including the public network, social media, major Internet companies, scientific institutions, management departments, and various enterprises, as well as data generated by individuals, such as photos, videos, and microblogs.

The arrival of the era of big data has attracted the attention of all countries, and a large number of big data RESEARCH and development programs have been launched. In order to promote the development of big data technology, both Nature magazine and Science magazine have successively launched special issues on big data. In 2012, the United States launched the Big Data Program and released the Big Data Research and Development Program, investing nearly 300 million DOLLARS in the research of big data technology[2].

Guo Lin[3]A traffic information fusion prediction model based on radial basis function neural network is established.Chen[4]Et designed a traffic flow prediction system based on MapReduce calculation framework;Cai Changjun[5]The actual traffic flow data are used to verify the advantages of the short-term traffic flow prediction model based on BP neural network in computing speed and adaptability.In the environment of explosive growth of data, if the data can be analyzed to find out the changes and further discover the potential information, then the future changes of traffic flow can be predicted, which can provide valuable reference for traffic flow prediction, traffic flow distribution, traffic accident prediction, road planning and so on.In addition, it can improve the ability to deal with emergencies. However, at present, the traditional information technology can no longer handle the massive traffic data, which also brings crisis to the intelligent transportation system.

ISSN: 2664-9640

2. Research on short-term Traffic Flow Prediction based on improved exponential smoothing algorithm in Hadoop environment

2.1. Exponential smoothing method

The exponential smoothing method is a kind of time series model, which makes smooth calculation on historical data, knows the changing direction of data, and then predicts the future situation. The advantages of exponential smoothing method include low modeling difficulty, fast operation and easy implementation, and it is suitable for processing large amounts of data.

Although the traditional exponential smoothing model is simple and easy to operate, it also has great shortcomings, which are mainly reflected in the lack of timely data prediction and the lack of ability to identify data turning points.This is because the smoothing coefficient selected by the model is fixed and it is difficult to obtain a coefficient consistent with the changing trend through the initial data.

In addition, the selection of groups of historical data is also a parameter that needs to be considered in the calculation. In the process of data acquisition, when the investigation time is long and the amount of data is large, which data is selected as the initial data for prediction has a great impact on the prediction results. Moreover, the solidification of the initial data limits the adaptability and flexibility of the model. In order to solve this problem, a method of updating smoothing coefficient and calculating range is proposed to improve the exponential smoothing model.

2.2. Dynamic exponential smoothing algorithm

In terms of traffic flow, if the range of data is relatively large, the trend of data change will also be constantly changing. In order to meet the requirement of traffic flow prediction, the smoothness coefficient changes constantly in the process of calculation.

The trend of traffic flow is stable within a short period of time, but due to various factors, the duration is changing constantly. If in the process of prediction, according to the effect of prediction, the selection range of data, namely the most appropriate calculation range, can be constantly updated while changing the smoothing coefficient, then the whole prediction process can achieve dynamic optimization.

The base of the improved prediction algorithm is exponential smoothing. Because the collected data is nonlinear, the method of cubic exponential smoothing is chosen in this paper. The model takes the square of the minimum error as the objective function.

xi was defined as the predicted value of group i, and yi as the measured value of group i. The current group number is M, n is the group number of measured data included in each calculation, and the maximum value of n defined is N. The current measured data obtained is defined as group Q.

Each time the algorithm performs a calculation, it produces a new set of optimal solutions. After the calculation is completed, the latest optimal solution (θ*,n*), and then calculate the optimal solution for the new set.

Algorithm flow:

Step1: Define the objective function as:

$$\begin{cases} f^*(\theta, n) = \min\ (x_i - y_i)^2 \\ 0 \le \theta \le 1 \\ 1 \le n \le \min(N, Q) \end{cases} \tag{1}$$

Step2: For each calculation, the change in θ value is 0.001, and the change in calculated group n is -1.

ISSN: 2664-9640

Step3: So let's start with the optimal value of θ, and when we're done with the optimal value of θ *, let's compute the optimal value of n*.

Step4: Initial value (0, min (N, Q)).

Step5: The predicted value corresponding to the optimal value θ* is x', and the corresponding calculated group number n is min (n, Q).

Step6: Each time n is calculated by continuously decreasing 1, the optimal predicted value x* is obtained, and the corresponding optimal solution is (θ*, n*).

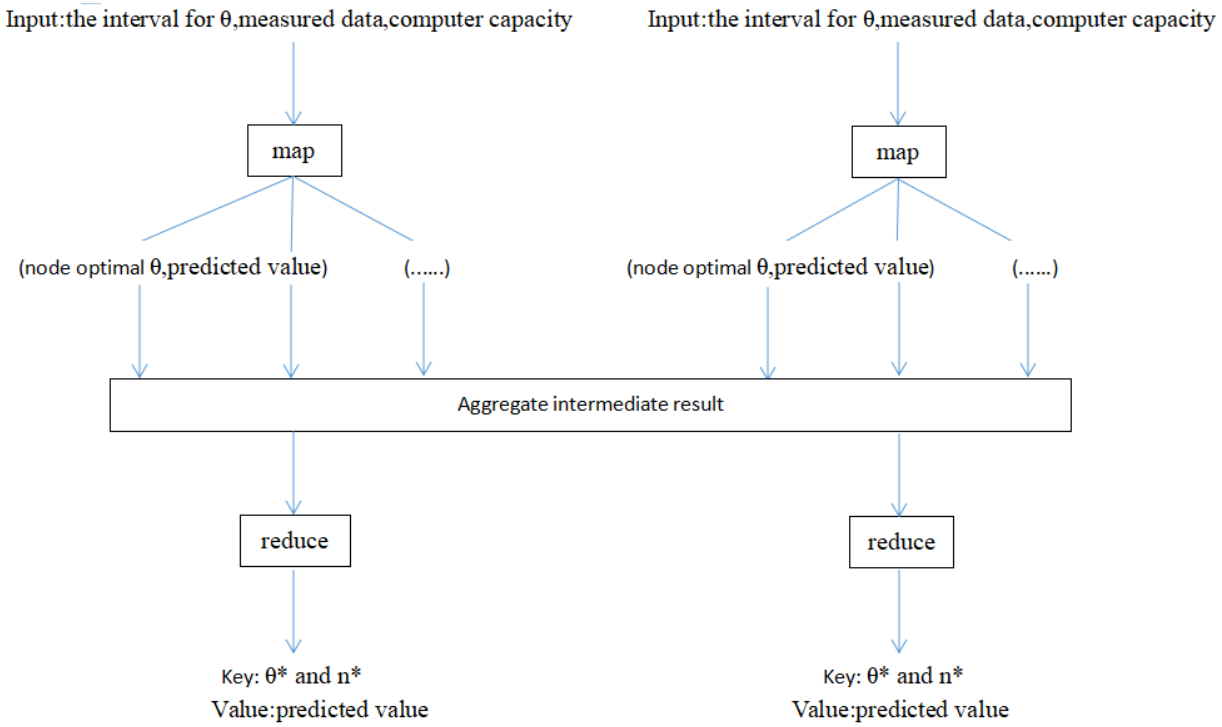

Figure 4 : Algorithm flow chart of dynamic shortest path based on Hadoop MapReduce

2.3. Algorithm model based on MapReduce

One of the bottlenecks to realize real-time traffic flow estimation is the processing performance of the algorithm for mass data. This paper introduces the MapReduce computing model in big data technology and parallelizes the induction algorithm.

(1) Data input

The key value is the theta value assigned, and the value is the measured raw data corresponding to the section. After Hadoop breaks up data, it divides them into multiple maps.The input data structure is shown in Figure 1.

Value of θ	calculation	Measured raw data

Figure 1 : Input data structure

(2) "Map" stage

The Map function takes the theta values, measured data and calculated range data as inputs, and calculates them according to the cubic exponential smoothing method. The Map function calculates the predicted value based on the theta value allocated and the calculated range to get the key and value values. Key is the optimal value of the node's allocated range, and value is the predicted value corresponding to the optimal value of.The output format of the "Map" phase is shown in Figure 2.

The node has the optimal theta value	Optimal node prediction value

Figure 2 : Output format of Map phase

(3) "Reduce" stage

Hadoop passes the results of the Map phase to the Reduce function. The Reduce function compares all predicted values to get the global optimal value and the corresponding optimal theta value. Using the optimal * value, all calculation ranges were traversed to obtain the optimal calculation range n* and the optimal predicted value.The output format of the "Reduce" phase is shown in Figure 3.

Global optimal (θ*, n*)	Global optimal predictive value

Figure 3 : Output format in Reduce phase

The data flow chart of the algorithm is shown in Figure 4.

3. Data Analysis

This paper takes Xingzhong Road, the most important north-south corridor in Zhongshan city, as a case study to verify the tracking and prediction ability of the model to the actual data.

The average speed of the road from the municipal government to Chenggui Road, which contains six red street lights, is obtained through the traffic detection system. The data collection time was from 7 to 20, a total of 13 hours. The data was set in 10 minutes, and a total of 78 measured samples were collected, as shown in Figure 5.

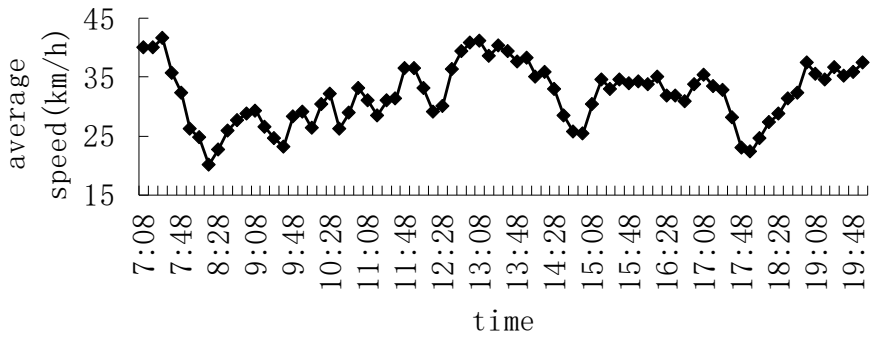

Figure 5 : Measured data

3.1. Exponential smoothing method

Because it is difficult to determine how many smoothing and smoothing coefficients should be used in the exponential smoothing method, this paper selects the first smoothing and the third smoothing. The smoothing coefficients are 0.3 and 0.7 respectively, and four experiments are conducted respectively.

When the smoothing coefficient is 0.3, the comparison between the predicted value and the measured value of the first exponential smoothing and the third exponential smoothing from 9:08 to 19:58 is shown in Figure 6.

When the smoothing coefficient is 0.7, the comparison between the predicted value and the measured value of the first exponential smoothing and the third exponential smoothing from 9:08 to 19:58 is shown in Figure 7.

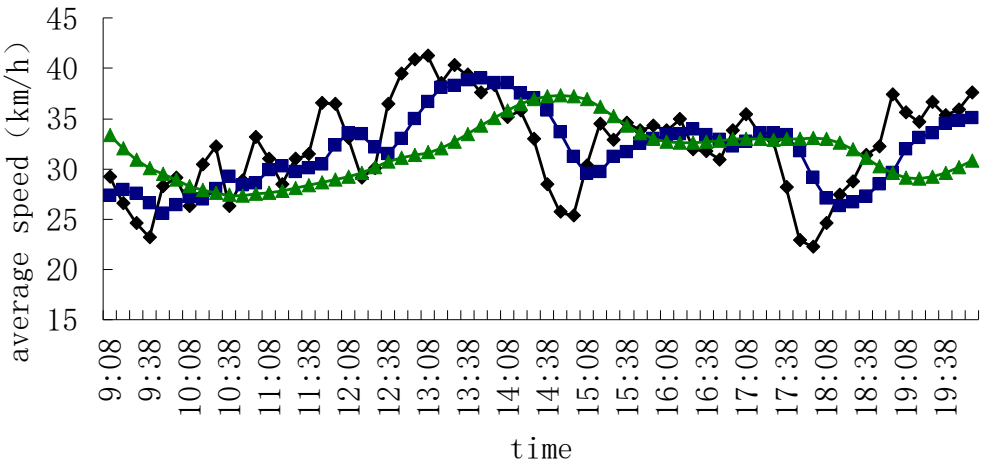

Figure 6: Comparison of the exponential smoothing predicted and measured values with a smoothing coefficient of 0.3

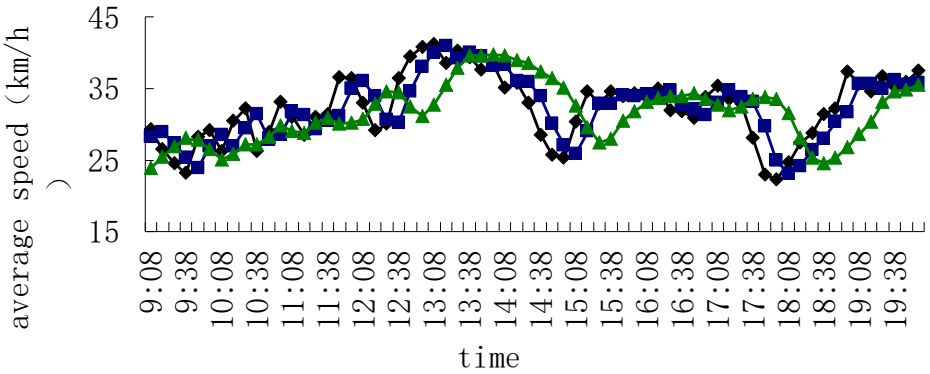

Figure 7: Comparison between the exponential smoothing predicted and measured values with a smoothing coefficient of 0.7

The evaluation indexes of the above four experiments are shown in Table 1:

Table 1: Results evaluation of exponential smoothing method

The evaluation index	θ= 0.3		θ= 0.7	
	First exponential smoothing	Cubic exponential smoothing	First exponential smoothing	Cubic exponential smoothing
Maximum absolute error	8.8343	11.9206	6.7075	11.2232
Mean absolute error	2.9045	4.2868	2.4073	3.9302
Mean absolute percentage error	9.42%	13.79%	7.83%	12.72%
Mean square error (mse)	12.8504	28.9131	8.7359	24.5758

According to the comparison in Table 4-1, the maximum absolute error minimum value appears in the primary numerical smoothing with a smoothing coefficient of 0.7. The minimum average absolute error value of 2.4073 and the minimum average absolute percentage error value of 7.83% both occur in the case of the primary exponential smoothing with the smoothing coefficient of 0.7, and there is not much difference with the index value of the primary

exponential smoothing with the smoothing coefficient of 0.3. In addition, the minimum value of the MEAN square error also appears in the first numerical smoothing with a smoothing coefficient of 0.7.

From the above analysis, it can be concluded that the primary exponential smoothing with a smoothing coefficient of 0.7 has an obvious effect on the recognition ability under abnormal conditions. This is due to the choice of a large smoothing factor, which enables rapid response to changes in observed values.

From the perspective of the whole model, whether the smoothing coefficient is 0.3 or 0.7, the prediction effect of the first exponential smoothing method is better than that of the third exponential smoothing method. The cubic exponential smoothing model, which USES the four indicators shown in the table as the evaluation criteria, fails to show the accuracy that the model should have. This is because the cubic exponential smoothing performs three calculations on the initial data, making the model less sensitive to the data. Therefore, after three times weighting the initial data, the model with three times exponential smoothing cannot react quickly to the changes of the data. Therefore, in the process of short-term traffic flow prediction, the advantage of cubic exponential smoothing is limited by the model.

3.2. Dynamic model

In the experiment to verify the dynamic model, the first calculated group number of the measured data obtained is Q=12, the change amount of each calculated smoothing coefficient is 0.001, and the change amount of calculated group number N is -1.The maximum number of sets of measured data calculated each time is N=78.

The comparison between experimental data and measured values is shown in Figure 8.

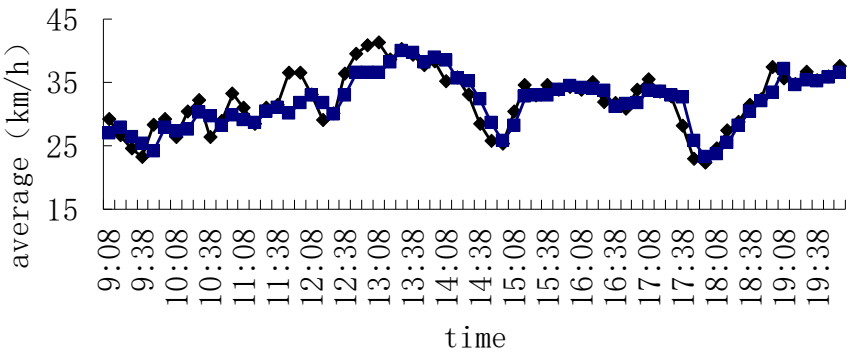

Figure 8: Comparison of predicted value and measured value

As can be seen from the figure above, the predicted results of the dynamic model are close to the measured values, and the prediction effect is better than that of the cubic exponential smoothing method.

The evaluation indexes of the dynamic model are shown in Table 2.

Table 2: Results evaluation of the dynamic model

The evaluation index	Maximum absolute error	Mean maximum error	Mean absolute percentage error	Mean square error (mse)
The dynamic model	6.3856	1.6540	5.23%	4.9583

It can be seen from the above table that the four index values obtained by using the dynamic model to process the initial data are all smaller than those obtained by using the exponential smoothing method. The maximum absolute error can reflect the model to recognize the fluctuation of data points. From the point of view of the average speed of the road, the average

absolute error is within the acceptable range, which shows that the prediction error of the model is small. The average absolute percentage error indicates that the prediction accuracy of the new model can reach 94.77%, which meets the requirements of the actual engineering accuracy level. The mean square error is small, which shows that the prediction results of the model change around the measured value with small error.

Experimental data show that in the short-term traffic flow prediction, the dynamic model can solve the shortcomings of the exponential smoothing model in sensitivity and delay, so that the prediction results are more accurate and meet the engineering requirements.

In order to better represent the relationship between the predicted value and the smoothing coefficient, this paper presents the change of the smoothing coefficient in the form of percentage, and the change process of the three is shown in Figure 9.

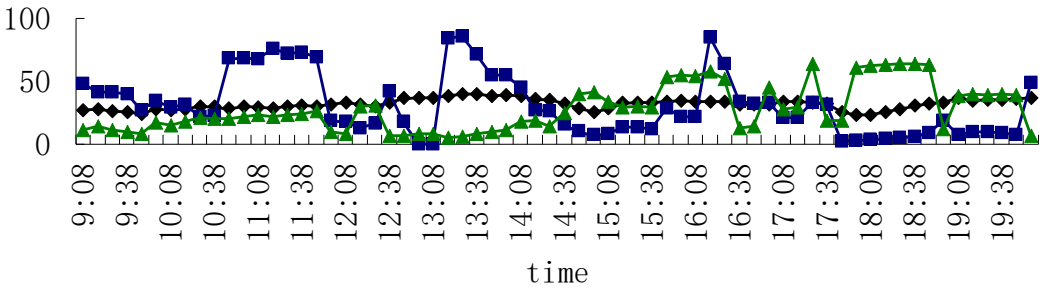

Figure 9: Model parameter analysis

It can be seen from Figure 4-6 that the smoothing coefficient and calculation range are constantly updated, forming a constantly changing predicted value. If one of the two variables, smoothing coefficient and calculation range, is changed, it is difficult to influence the predicted value. For example, the two data points changed greatly from 10:38 to 10:48, the smoothness coefficient increased by 0.4565, the number of calculated groups did not change, while the predicted value changed by only 1.5117km/h. Therefore, only when the smoothness coefficient and the calculated group number change together, can the prediction result change with the trend of the measured value, and then change the deficiency of the cubic index.

Pearson correlation analysis was used to analyze the smoothness coefficient and calculate the number of groups. The results showed that there was a significant negative correlation between the two (P=0.006) and the correlation coefficient R was -0.337. If the smoothing factor decreases, the number of calculated groups will increase, and vice versa. The more the smoothing factor approaches 0, the more consistent the model is with the trend of earlier data. On the contrary, the more the smoothing coefficient tends to 1, the more similar the trend of the model is to that of the recent data. Therefore, the reduction of the number of data calculation sets can improve the effect of the recent data on the prediction results.

4. Conclusion

Intelligent transportation system plays a key role in solving congestion problems, and traffic flow prediction is an important part of intelligent guidance of intelligent transportation system. In this paper, the traditional exponential smoothing model is optimized, and a parallel computing model based on MapReduce is proposed in the case of massive traffic flow data.

Through the parallelization calculation based on MapReduce framework, the computational efficiency of the prediction algorithm in this paper is greatly improved. This algorithm can calculate the massive traffic flow data of large-scale road network, and can realize real-time traffic flow prediction, and provide reliable basis for traffic flow prediction, traffic flow allocation and scientific path planning.

ISSN: 2664-9640

References

[1] Ni Ning. Exploration and Research of E-commerce Platform in the Era of Big Data -- Taking Taobao as an example [J].Jiangsu Business Theory, 2014(5) : 13-14.

[2] Wang Yibo, GUO Xin, WANG Jimin.Big Data Research Topic Analysis based on Word Co-occurrence [J]. Library Forum, 2014(8) : 96-102.

[3] Guo Lin. Research on Traffic Information Collection based on Information fusion [D]. University of Science and Technology of China, 2007.

[4] Chen C, Liu Z, Lin W H,et al. Distributed Modeling in a MapReduce Framework for data-driven Traffic FlowForecasting[J].IEEE Transactions on Intelligent Transportation Systems, 2013, 14(1):22-33.

[5] CAI Changjun. Research on BP Neural Network Short-term Traffic Flow Prediction Model [J]. Fujian Computer, 2015(3).