Mining and Analysis Based on Big Data in Public Transportation

Yinxin Bao[1], Chengyu Zhang[1], and Quan Shi[1,2](✉)

[1] School of Information Science and Technology,
Nantong University, Nantong, China
sq@ntu.edu.cn
[2] School of Traffic, Nantong University, Nantong, China

Abstract. City bus is an important part of the city transportation system. Reasonable and healthy bus line can greatly improve the city's economy and environment. This paper is based on mass data generated by the public transportation system of Nantong, such as GPS data of buses and passengers' swiping card data. Through the cleaning, matching, calculation and mining of the original bus data, this paper can provide decision-making basis for bus companies in terms of transfer rate, non-straight line coefficient, scheduling conflict rate, station accessibility, station coverage and passenger number. The LSTM model is used to predict the arrival time of the bus, and the accuracy of the model is improved by influencing factors such as weather change and direction.

Keywords: Big data in public transportation · Bus operation indicator · LSTM · Data visualization

1 Introduction

Urban public transport is an important part of urban public facilities. A mature and reasonable public transport network directly affects the economic development and construction of the city. In the context of the era of big data and Internet of Vehicles, expanding scale of advanced intelligent and informational public transport system construction, along with the breadth and depth of data collection has increased dramatically. The intelligent bus subsystem has accumulated massive traffic data for the development of smart traffic [1]. The paper selects urban buses as the research object for in-depth analysis and research. According to the data generated for each submodule of the intelligent public transport system, analysis of bus operation time, number of passengers, regional bus accessibility, site accessibility, travel demand and other operation status and characteristics. Provide decision-making basis for bus operation analysis.

K. Li et al. (Eds.): ISICA 2019, CCIS 1205, pp. 681–688, 2020.
https://doi.org/10.1007/978-981-15-5577-0_54

2 Overview of Related Technologies

2.1 Development Technology

The data processing techniques used in this paper mainly include: Hadoop, HDFS, HBase, MapReduce, Spark.

Hadoop is a distributed system infrastructure developed by the Apache Foundation [2]. It is suitable for data storage and computation on multi-computer cluster. The two most important parts are MapReduce Distributed Computing System and HDFS Distributed File System [3]. Hadoop provides a secure and efficient access service and a secure and reliable large data processing engine for massive data processing.

HBase mainly solves the problem of high-speed storage of large data and reading distributed database.

HDFS is the basis of data storage management in Hadoop system. It is a high fault-tolerant storage system. It can not only detect and deal with data loss caused by hardware failures, but also store and share data of each computing node across multiple computers.

MapReduce is a parallel computing model between a large number of data sets, which is mainly used for large-scale data computing. MapReduce mainly consists of two parts: Map and Reduce. 'Map' is mapping data node, and 'Reduce' means simplifying data.

Spark is a framework for data reading and data computing through computer memory. In terms of speed, Spark uses and extends MapReduce computing model to achieve a universal and efficient computing platform [4]. Spark supports more computing models, including flow processing and interactive queries.

2.2 Development Tool

The development tools used in this paper mainly include: Flask back-end framework, Vue front-end framework and TensorFlow framework.

Flask is a back-end framework written in Python language. It uses Werkzeug's WSGI toolbox and Jinja2 as its template engine. Vue is a progressive framework for building user interfaces [5]. It is a layer-by-layer application from the bottom up. The core of Vue lies in the layer, which has a wide range of third-party library resources and project integration.

TensorFlow is an open source software library that uses data flow diagrams for numerical computation [6]. TensorFlow is an open source library developed by the Google team. It can be built on multiple platforms because of its flexible and expansive architecture through the calculation of deep neural network.

3 Analysis of Bus Operation Characteristics

3.1 Raw Data Manage

The data used in this paper are mainly provided by Nantong Public Transport Company. The original data are divided into two parts: dynamic data and static data, as shown in Table 1. The dynamic data are the resident's bus card swipe data, bus dispatch data, bus passenger flow counter data, etc. The static data are mainly the location data of bus stations, bus line data and so on.

Table 1. Types of original bus data.

Data type	Data	Data description
Static data	Bus station data	Site coordinates, Site names, Site numbers
	Bus line data	Line number, Starting station, Terminal station, Passing station
Dynamic data	Bus IC card	IC card number, IC card type, Line number, Transaction time
	Bus dispatch data	Departure time, Line number
	Bus location data	GPS Data

Static Data Processing. The static data used in this paper are mainly bus station data and bus route network data. Bus stop data mainly includes the site coordinates, site number, site name and so on. Bus route network data mainly includes starting station information, terminal information and route number. This paper integrates two kinds of static data and integrates the GPS data of the station into the bus line data, which can display the road network information of the bus system visually on the map [7].

Dynamic Data Processing. The dynamic data used in this paper are mainly divided into IC card data and bus GPS data. The original data mainly come from bus card swipe data provided by bus companies and GPS data during bus driving.

The IC card data and bus GPS data are processed respectively.

1. Data cleaning. Clean the bus IC card data and bus GPS data to remove invalid or incorrect data from IC card data and GPS data.
2. Data preprocessing. The time variables in the data are cut and spliced into time stamps, and the GPS data are converted into coordinates under the GCJ-09 standard.
3. Data matching and integration. The collected coordinate data are mapped and the GPS data deviating from the road are eliminated. Some data of original IC card swipe data contain a large number of data fields, some of which are not analyzed and studied in this paper. These redundant data need to be integrated and deleted in order to reduce the waste of time and memory space in subsequent data processing and calculation.

3.2 Bus Operation Indicators

The bus operation indicators analyzed in this paper mainly include: bus transfer rate, non-linear coefficient, dispatch conflict, site accessibility, site coverage and IC card data statistics [8].

As one of the important indicators to evaluate the quality of public transport service, the transfer rate of public transport needs to be analyzed. Transfer data of Nantong passengers are shown in Fig. 1.

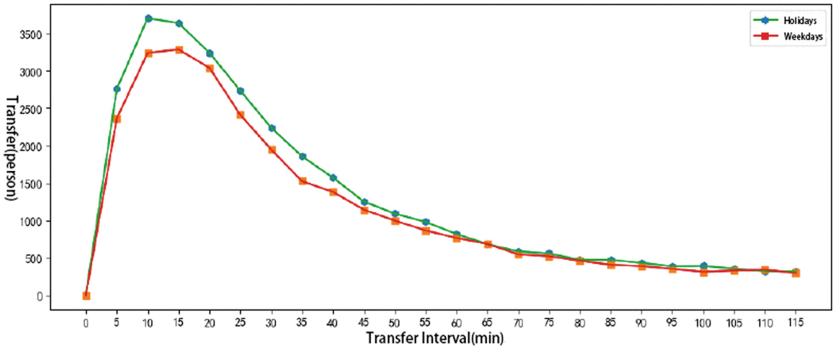

Fig. 1. Public switching multiplier data

Non-linear coefficient is an important index in the layout planning of public transport network [9]. The larger the non-linear coefficient is, the longer the detour distance of the line is, the lower the comfort of the residents. Excessive non-linear coefficient will lead to long ride time and uneven local passenger flow, while too low non-linear coefficient will lead to inconvenience for residents to transfer. The non-linear coefficient of Nantong's lines is shown in Fig. 2.

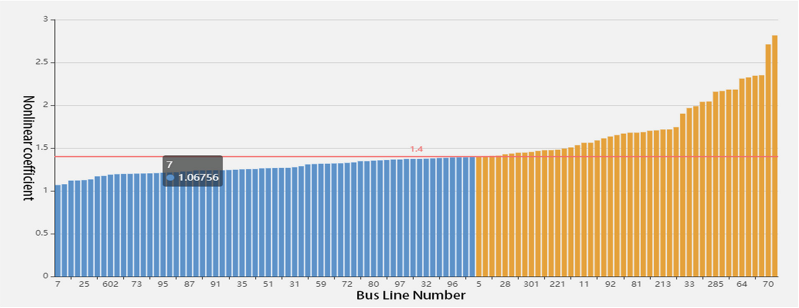

Fig. 2. Non-linear coefficients

Bus scheduling relies on fixed time intervals for departure. Because of road condition changes, weather factors and other factors, two buses on the same line will pass through the same station one after another. When two or more buses pass the same station within five minutes, dispatching conflicts are considered to occur, as shown in Fig. 3.

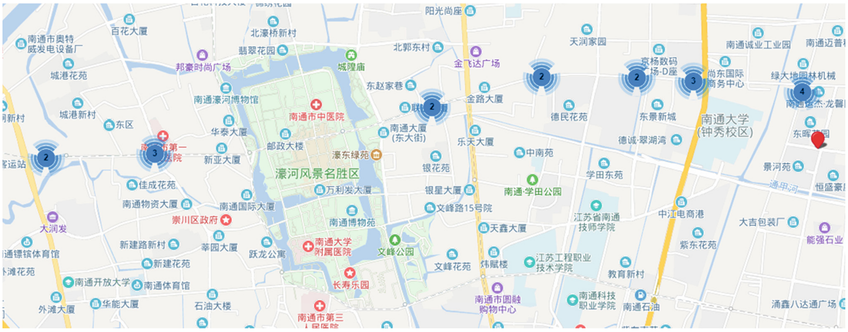

Fig. 3. Bus dispatch conflict

As an important part of analyzing the service quality standard of urban public transport, the accessibility of bus stops affects the travel efficiency of residents, and is also an important basis for residents to choose the way of travel.

Make statistics of different types of bus card swiping data, that is, analyze the total swiping data of one day, count the number of different time periods and different types of CARDS, and understand the characteristics of different groups.

Bus station coverage rate is an important index to evaluate the service quality of public transportation system in this region. The coverage of bus stops directly affects the convenience of residents to travel. According to calculation, the coverage rate of 300 meters service area of Nantong bus station is about 64.5%, which is 50% higher than the national standard in GB T 22484-2016.

4 Bus Arrival Prediction Model

4.1 Model Selection

The prediction model used in this paper is LSTM, which is a variant of RNN. Compared with RNN, LSTM solves the problem of long-term dependence of RNN [10]. RNN is applied to current tasks by connecting the factors that influence the previous event information. However, when predicting data sets with large time intervals, the gradient will disappear due to the increase of dependence factors. LSTM solves the problem of long-term dependence by selective forgetting and long-term memory of processed information.

4.2 Data Preprocessing

In order to improve the computational efficiency and fitting effect of the algorithm, it is necessary to remove variables that have no or minimal impact on the prediction results before data prediction by LSTM model. This paper uses Lasso algorithm to judge the weight of all variables in the original data. The calculation method is shown in formula (1).

$$\breve{\beta}^{lasso} = min\left\{\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(y_i - \beta_0 - x_i^T\beta\right)^2\right\} \tag{1}$$

In formula (1), x is the training data matrix, β is the regression coefficient, and y is the training label. The formation of training data sets is shown in Table 2.

Table 2. Training data table

Attribute	Explain
UPDAWN	Up and down
WEATHER	Weather condition
STARTTIME	Departure time
DISTANCE	Site distance
WEEKDAY	Date
STOP	Site number
BUSNO	Vehicle number
STOPTIME	Arrival time

4.3 LSTM Model

LSTM is a chain-structured cyclic neural network model, which predicts the arrival time of the bus by calculating the forgetting gate, input gate and output gate. Its structure is shown in Fig. 4.

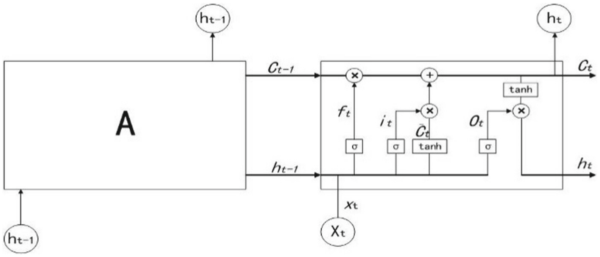

Fig. 4. LSTM structure

After data input, the first step is through the forgetting gate, which determines the current cell data input and the information discarded in the last cell state. The forgetting gate reads h_{t-1} and x_{t-1} and outputs a value between 0 and 1 through formula (2) and updates it to the state variables of C_{t-1} cells.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \tag{2}$$

In formula (2), σ denotes the activation function Sigmoid. The definition of Sigmoid function is shown in formula (2). W_f represents the weight of forgetting gate and b_f represents the bias matrix of forgetting gate.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

In formula (3), x represents the input variable, and the value of the σ function is between 0 and 1.

The updated information is determined by formula (4) and formula (5). Through the Sigmoid function which data is worth updating, through the tanh function to create a new candidate value vector.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\check{C} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{5}$$

In formula (4) and formula (5), Wi represents the input gate weight, bi refers to the input gate bias matrix, WC refers to the weight matrix of tanh function, b_C refers to the bias matrix of tanh function, and Č is the candidate cell state.

Formula (6) is used to determine the cell status of the current cell.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \check{C}_t \tag{6}$$

The output information is determined by formula (7) and formula (8).

$$O_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = O_t \cdot tanh(C_t) \tag{8}$$

4.4 Fitting Results

The fitting rate of LSTM model is about 93.5921%.

5 Summary

This paper implements a data mining and analysis system based on urban bus data, which is mainly divided into four parts: data storage, data analysis and calculation, data prediction and data visualization. The data storage in the system is distributed through Spark framework. Data processing is mainly to sort out the original data and calculate the bus operation indexes such as the bus line's non-linear coefficient, passenger transfer rate, bus station coverage rate, etc. Data prediction is mainly based on LSTM model to predict bus arrival time. Data visualization is achieved by using a front-end framework (Vue) and a back-end framework (Flask).

Acknowledgments. The National Natural Science Foundation of China (61771265), The Science and technology Project of Nantong (CP12017001).

References

1. Zhou, H.: Research on traffic congestion management based on big data rule mining [1]. Stat. Inf. BBS **32**(05), 96–101 (2017)
2. Feng, X., Wang, W.: Research on Hadoop and Spark application scenarios [1]. Comput. Appl. Res. **35**(09), 2561–2566 (2018)
3. Hao, Y.: Design of CDH-based data visualization platform [2]. Chengdu University of Technology (2018)
4. Liu, P.: Design and implementation of data management platform based on Spark [2]. Zhejiang University (2016)
5. Wang, Z.: Design and implementation of development platform based on Vue.js [2]. Guangdong University of Technology (2018)
6. Wang, Q., Liang, J.: Research on traffic sign recognition based on TensorFlow [1]. Value Eng. **38**(27), 204–206 (2019)
7. Li, M.: Public transportation intelligent scheduling and location publishing system based on GPS [1]. Comput. Prod. Circ. **10**, 127–128 (2018)
8. Chen, J., Lv, Y., Cui, M.: Judgment method of passenger disembarkation station with IC card based on travel mode [1]. J. Xi'an Univ. Arch. Technol. (Nat. Sci. Ed.) **50**(01), 23–29 (2018)
9. Zhao, S., Lu, Y., Hao, L.: Intercity bus route planning based on node importance in urban agglomeration [1]. Highw. Mot. Transp. (05), 24–26+30 (2018)
10. Yang, Q., Wang, C.: Prediction of global stock index based on deep learning LSTM neural network [1]. Stat. Res. **36**(03), 65–77 (2019)