# Mining and Analysis of Big Data Based on New Energy Public Transit

Yinxin Bao[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.** With the rapid development of new energy vehicles, urban public transport vehicles are gradually dominated by new energy public vehicles. This article collects and analyzes various data in the system of new energy bus, and mines the basic information needed for the Intelligent bus system. It also analyzes and integrates related data, explores the operation characteristics and laws of new energy buses, and establishes a new energy bus data mining and analysis system to provide important decision-making basis for the operation and management of new energy buses. This paper combines big data processing technology with visualization technology to realize the new energy public transportation data visualization system. This system can provide bus operators with the basis for optimization of public transportation network, which greatly improves the management level of public transportation enterprises and the operating efficiency of public transportation vehicles.

**Keywords:** Big data of new energy bus · Operational analysis · Data mining · Data visualization

## 1 Introduction

Nowadays, sustainable development has been strongly advocated by the government, urban public transport vehicles are gradually transitioning from fuel vehicles to new energy vehicles. The advantages of new energy public vehicles are low pollution, high utilization rate, low noise, convenient maintenance and high safety performance, which is the trend of public transportation development in the future [1]. New energy buses have abundant data resources. As long as relevant data are used and data collection and analysis are done well, the basic information needed by the intelligent bus system can be mined. By analyzing the data information from different sources and integrating them organically, the characteristic values and rules related to bus operation can be found out, which can provide important decision-making basis for the operation and management of new energy bus system.

K. Li et al. (Eds.): ISICA 2019, CCIS 1205, pp. 739–746, 2020.
https://doi.org/10.1007/978-981-15-5577-0_59

# 2 Data Acquisition and Data Preprocessing

## 2.1 New Energy Bus Data Collection

At present, data collection technologies commonly used on buses include: bus card swiping device, bus track positioning device, passenger flow counter, etc. Data collection methods are as follows.

1) **New energy bus credit card data collection method**
   At present, the one-vote system is more commonly used in China. The passengers complete the card swiping through the handheld IC card on the bus, and each time the card is swiped, the passenger swipe data is generated at the card swiping machine terminal [2, 3].
2) **New energy bus trajectory data collection method**
   By carrying a positioning system, the new energy bus can send real-time information such as vehicle position coordinates, speed, running state and running direction to the dispatching center [4].
3) **New energy bus passenger flow data collection method**
   The new energy buses put into operation in Nantong are equipped with passenger flow counters, which can collect the arrival time of the bus and the data of the upper and lower passenger flow at the station [5].

## 2.2 New Energy Bus Data Preprocessing

This section mainly preprocesses the data of passenger swipe card data, bus track data, passenger flow counter data, station and line data of new energy bus.

The spark SQL in spark is mainly used for data preprocessing of passenger swiping card data to eliminate the error location in the original data. The data preprocessing of the bus location data is mainly based on the location of the bus, considering that the bus runs in the urban area of Nantong City, so the data out of this range are eliminated [6]. Because the passengers on the bus will walk back and forth, resulting in repeated counting of the passenger flow counter, considering that the standard number of passengers in Nantong city is 35 to 43, so the data of the number of passengers exceeding the standard number is eliminated. Analyze and process station, line number, up and down lines, station sequence and positioning data. Through data analysis, this kind of data mainly exists the situation of latitude data missing.

# 3 New Energy Bus Data Analysis and Mining

## 3.1 Data Analysis

### 1) Analysis of passenger flow data

Nantong City passenger flow analysis is mainly aimed at the analysis of bus card data and passenger flow counter data. Firstly, the bus card data is analyzed according to the line, operation time and bus card data statistics. The card swiping data of the public transport card is counted according to the card swiping personnel and the public transport operation line, among which the card swiping personnel are divided into three types: student card, senior card and other cards. Because of the prepossessing of the original card swiping data, the original card swiping data is divided into IC card, M1 card and citizen card. As shown in Fig. 1, spark SQL is used to analyze the processed data, and python calls Matplotlib library to display the card reading data [7, 8].

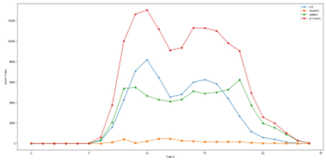


**Fig. 1.** Swipe statistics

### 2) Trajectory data analysis

Due to the complexity of the road conditions, the bus arrival time will deviate from the original planned arrival time. When this happens, it will lead to the aggregation of buses departing at different times in the line and reduce the efficiency of the bus operation line [9]. In this paper, the GPS track data of new energy bus is used to analyze the phenomenon of bus aggregation, and a reasonable model is established. The aggregation calculation of new energy bus in Nantong is carried out. Figure 2 shows the traffic flow on a line in Nantong on October 1, 2018. The more dense the number of points, the more bus aggregation occurs.

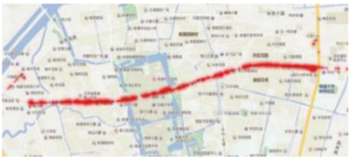


**Fig. 2.** A line conflict situation on October 1, 2018

3) **Line network analysis**

In order to better plan the operation scheme of Nantong new energy bus, this paper makes a comprehensive evaluation of the existing network and routes of Nantong city, and selects three indexes of site coverage rate, site density and non-linear coefficient to analyze.

Bus station coverage rate is an important basis to judge whether the allocation of regional public transport resources is reasonable. In general, site coverage is defined as the coverage radius of a bus stop within a specified area of 300 m or 500 m. The coverage area can be obtained by calculating the total coverage area of the site divided by the total area of the specified area. Nantong bus station information table data has 1660, this paper calculates the site coverage of each site. The result shows that the service coverage of 300 m in Nantong bus station is about 64.5%, which is much higher than 50% of the national public tram passenger service standard.

The site density is the ratio of the total number of stations parked by the bus to the total area of the city. This feature reflects the distribution of the bus station under the current area and its adaptation. Using Nantong urban station site data, statistics were made on all bus stations in Nantong City, and the density of Nantong bus stations reached 2.295.

The nonlinear coefficient is the ratio of the actual distance of a line to the first and last distance of a line. The larger the nonlinear coefficient, the larger the service area of the line. The smaller the coefficient is, the smaller the service area of the line is. If the service area is too small, it can be solved by adding bus lines.

## 3.2 Data Mining

Accurate prediction of bus arrival time can not only significantly improve the operational efficiency and service level of public transportation, but also significantly promote the development of public transportation. This chapter will study the arrival time prediction of Nantong New Energy Bus and explore how to establish a prediction model for the arrival time of new energy buses.

1) **Model establishment**

Based on the study of RNN recurrent neural network, this paper establishes the GRU neural network model to complete the prediction of bus arrival time. Different from the traditional LSTM, the GRU model only has two gates, which do not control and retain internal memory. Therefore, the model is simpler and more efficient than LSTM. Figure 3 shows the GRU network model diagram used in this paper [10].

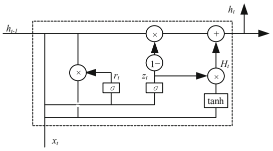


**Fig. 3.** GRU network model diagram

2) **Model calculation**

Data of different categories need to be processed by different standardized methods. For fixed types of data defined by category labels such as car number, week, and weather, zero mean method can be used for standardized processing. The formula is shown in Eq. (1).

$$m^* = \frac{(x - \mu)}{\sigma} \tag{1}$$

In Eq. (1), $m*$ represents the normalized data, $x$ represents the pre-processed data, $\mu$ represents the sample mean with characteristics, and $\sigma$ represents the standard deviation of the sample.

In this project, Lasso algorithm is mainly used to filter the original eigenvalues of new energy bus, as shown in Eq. (2).

$$\widehat{L}_{Lasso} = \text{argmin}_{\beta} \sum\nolimits_{i=1}^{n} (m_i - L_0 - \sum\nolimits_{j=1}^{p} x_{ij}\beta_j)^2 \tag{2}$$

In Eq. (2), $x_{ij}$ is the $j$-th variable of the $i$-th group of eigenvalues. Let row vector L be the regression coefficient in the eigenvalue and $m$ be the training label.

In Fig. 3, this paper analyzes the network element of the layer. When the time step is step t, the update gate calculates the data through the Eq. (3).

$$z_t = \sigma\left(W^{(z)}x_t + U^{(z)}h_{t-1}\right) \tag{3}$$

In Eq. (3), $x_t$ represents the $t$-th component of the input value $x$, $W^{(z)}$ represents the weight matrix, $h_{t-1}$ holds the information of the previous time step, and linearly transforms through the weight matrix $U^{(z)}$.

The purpose of the update gate is to add it and convert the result through the Sigmoid function to compress the activation result between 0 and 1. The update gate determines how much historical data is passed to the future, reducing the risk of gradients disappearing. The reset gate determines the forgetting process for the data and can be expressed by Eq. (4).

$$r_t = \sigma\left(W^{(r)}x_t + U^{(r)}h_{t-1}\right) \tag{4}$$

3) **Model training**

After model training, Fig. 4 shows the Loss function curve of GRU network model under Tensorboard environment. It can be seen that the model rapidly converges during training, and the Loss value function tends to be stable after the 10th cycle.

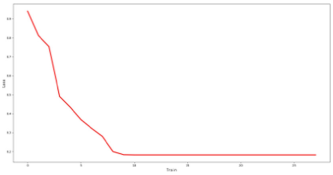


**Fig. 4.** Loss function curve

4) **Model fitting**

After the model training is completed, this paper compares four days of bus operation data with the predicted results. Figure 5 shows the fitting curve of GRU model prediction and real data. As shown in the figure, the fitting effect is relatively ideal. By using the linear regression fitting index r-square as the model evaluation index, the fitting degree of GRU neural network is 94.6138%.

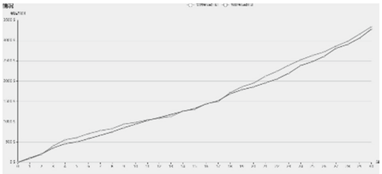
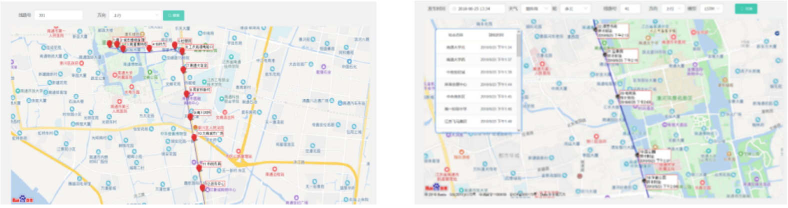


**Fig. 5.** GRU model prediction and real data fitting curve

## 4 Realization of New Energy Public Transport Visualization Platform

The core functions of the system are divided into the following parts: bus line visualization module, bus station visualization module, bus passenger flow visualization module and bus arrival station prediction module.

In the bus line visualization module, the line query function is designed, as shown in Fig. 6(a), to provide line query service for users. When the user wants to access a place, he can get whether the current line is accessible. The bus station visualization module realizes the visual query function of Nantong bus station. This module shows through the thermal map of the website that the dark sites are more densely distributed on the map, whereas the light sites are more loosely distributed on the map. Passenger flow visualization module is a visual analysis of bus passenger flow status. The system can divide passenger card data into elderly card, student card and ordinary card according to the type. This module can show the bus swiping card and the number of passengers in Nantong city. The bus arrival prediction module shown in Fig. 6(b) details the bus arrival prediction model based on GRU neural network. This system can accurately predict the bus arrival time.



(a) New energy bus visualization module

(b) Arrival prediction module

**Fig. 6.** Visual module

## 5 Conclusion

This paper aims to analyze the problems and deficiencies existing in the current public transportation system through the visual analysis and data mining of new energy public transportation data. On the basis of studying the theoretical basis of big data processing technology, data mining, analysis and processing, and data visualization, this paper analyzes the operation and route indicators of new energy buses by combining bus passenger swipe card data, bus network data, and bus positioning data. The visualization of new energy bus data mining system is realized to provide decision support for bus operators.

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

## References

1. Xingqiang, C., Ke, T., Yuechuan, C., et al.: Research on big data decision support platform for new energy public transport vehicles. Urban Public Transp. **12**(03), 25–29 (2017)
2. Qunyong, Z., Keyun, S., Zhijie, Z.: Time and space analysis of bus passenger flow based on massive ic card data. J. Guizhou Univ. (Nat. Sci.) **35**(06), 93–98+105 (2018)
3. Ting, Y., Guilian, F., Xinghui, Ma.: Bus passenger flow calculation based on bus IC card information. Traffic Eng. **18**(06), 51–56 (2018)
4. Andeng, Y.: Bus od calculation based on bus GPS and IC card data. Harbin Institute of Technology (2017)

5. Zhouquan, W.: Research on temporal and spatial distribution of bus passenger flow based on IC card data and GPS data. Southwest Jiaotong University (2016)
6. Andrienko, N., Andrienko, G., Rinzivillo, S.: Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics. Inf. Syst. **57**(02), 172–194 (2016)
7. Zhu, J.: Research on data mining of electric power system based on Hadoop cloud computing platform. Int. J. Comput. Appl. **41**(4), 22–24 (2019)
8. Chaolong, J., Hanning, W., Lili, W.: Research on visualization of multi-dimensional real-time traffic data stream based on cloud computing. Procedia Eng. **137**(8), 24–26 (2016)
9. Perveen, S., Kamruzzaman, M., Yigitcanlar, T.: What to assess to model the transport impacts of urban growth? A Delphi approach to examine the space–time suitability of transport indicators. Int. J. Sustain. Transp. **13**(8), 31–34 (2019)
10. Ke, K., Hongbin, S., Chengkang, Z., Brown, C.: Short-term electrical load forecasting method based on stacked auto-encoding and GRU neural network. Evol. Intell. **12**(3), 2445 (2019)