Analysis of Bus Line Passenger Flow Based on Big Data—A Case Study of Guangzhou Bus Line

Yingqi Ma¹ and Weichuan Yin²,*

¹School of Finance, Anhui University of Finance and Economics, China.
²One Belt-One Road Strategy Institute, Tsinghua University, China.

*Corresponding author

Abstract. The passenger flow analysis which is based on the public transport passenger travel data could provide an important reference for the urban transit network planning and layout optimization. 8,926,605 smart card records of Guangzhou bus lines from August to December in 2014 and weather data have already been collected. Ali cloud computing technology is used to process the big data and analyze passenger flow characteristics. The test between Ali cloud computing technology and traditional SQL database technology has been made. K-means algorithm is utilized to cluster the passenger travel frequency, and the structure of the passenger travel type is gained. The results show that the seasonal weather have some influence on passenger flow each month, and the group of older residents is higher than the group of students in travel frequency. The travel frequency of passengers in low level accounts for more than 80%. Cloud computing technology is used to deal with the such big data, which has a good adaptability.

Keywords: Urban transit; Passenger flow analysis; Big data; Smart card data; K-Means algorithm.

1. Introduction

In recent years, with the rapid development and application of computer and information technology, the public transportation field accumulates a large amount of passenger travel data and it has a great significance on mining value from the analysis on the big data. The big data technology has some great advantages in doing with these problems. With the decision-making demand become increasingly complex, the application scenarios of public transport data are more and more abundant[1]. Passenger flow analysis based on public transport passenger travel data that can provide an important reference for the urban bus transit network planning and layout optimization. The analysis of bus passenger flow characteristics based on collection of the smart card data can help to realize the reasonable resources allocation, and could better satisfy the traveling demand of passengers, which also stimulates the efficiency of public transport operation. What is more, it also has a vital research value to reduce the imbalance of supply and demand.

2. Literature Review

Researches on smart card data for passenger flow analysis are as follows. Chen and Liu[2] analyze the spatial and temporal characteristics of the passenger flow in Beijing with smart card data. Dai[3] makes some research on the relationship between the usage of public transport smart card and the characteristics of public transport. Shu[4] analyzes passenger flow rule from data mining method with
smart card data. Ma et al[5] mines the passenger travel behavior with the $K$-means clustering and rough set algorithm based on 3,000,000 Beijing smart card records in a week. Morency et al[6] analyze the passenger travel behavior based on 277 days of Canadian smart card records. In the research of big data about traffic mode of public transport, the source of basic data is varied. Xiao et al[7] analyze the change trend of the traffic in the 2008-2014 of China based on the data of the expressway network toll collection system.

There exists some defects about the passenger flow analysis research with big data as follows: (1) the weather and some other external factors have been taken into consideration in the passenger flow analysis research; (2) the big data analysis and processing technology still stay in the traditional database level and are not related to distributed cloud computing technology in the process of transportation research.

From this paper, we concentrated on the passenger flow analysis that is based on big data technology. The cloud computing technology has been used to process the big data of Guangzhou bus lines. The influence of the weather factor is also taken into accounts, and the characteristics of the public transit passenger flow and the fluctuation rules are analyzed.

3. Data description
The data is provided by the Guangdong Lingnantong incorporated company. The line names are fuzzy processed for confidentiality requirements. There are 5 months smart card data of Lingnantong passengers of Guangdong bus lines between August 2014 and December 2014, involved nearly 2 million passengers, 2 bus lines and 8,926,605 data records.

4. Methodology
An analysis method of bus big data is put forward in this paper, which included data reading, data storage, data cleaning and data analysis by these four steps. The data analysis step includes the scale, ratio of passenger flow, card type and clustering. The flow chart of this analysis method is shown in Figure 1.

![Flow chart of big data analysis method.](image_url)

5. Case study
A case study of Guangzhou bus line is made by Ali cloud ODPS[8] by the four analysis aspects. The Ali cloud platform is a computer cluster with some integrated data mining algorithms that is included in this paper, which has a good result of processing the bus line big data and have a good development prospect. The users needn’t know the inner core or mechanism of big data processing technology. The only thing...
to do is to collect the specific data set and analyze it from the professional views. They are namely passenger flow, types, areas of the passenger, weather factors impact, and cluster of passengers travel frequency.

5.1 Passenger Flow Analysis

Statistics of Guangzhou bus line big data includes: monthly passenger flow, passenger flow of bus line, the analysis of monthly maximum hourly cross section passenger flow, proportion relationship between bus vehicle and passenger flow, as are shown in Figure 2-3.

Passenger flow of bus line 10 is a little higher than line 15, and the monthly passenger flow fluctuation of each line is the same as passenger flow fluctuation. The passenger flow in August 2014 is the lowest. The student card records account for 4.95% of the total records in August 2014. The percentages of student card records are 5.64%, 5.80%, 6.21% and 6.08% in other four months respectively. The student card records accounts for 8.56%, 9.94%, 11.42% and 10.52% of increasing records between September and December compared to August in 2014. So, the summer holiday in August decreases students travel frequency since they do not need to go to school. While Guangzhou has hot weather in August, travel willingness of residents may also decrease, but it is unable to determine the inevitable connection between these two.

The monthly maximum section passenger flow (MSPF) is 7,866 records in total and appears at 8-9 am on August 5 2014, others appear at 8-9 am on September 23rd 2014, at 8-9 am on October 21st 2014, at 8-9 am on November 4th 2014 and at 8-9 am on December 10th 2014 respectively. The daily maximum section passenger flow appears between 8-9 am in each month, the possible reason is that 8-9 am is the peak commuting time. The peak point of daily maximum section passenger flow appears in summer season, hot weather decreases the passenger flow at noon and in the afternoon or emergent events happen that leads to a surge in passenger flow may be the reason. The monthly average maximum section passenger flow (AMSPF) reflects passenger flow of peak hour within a month. It shows that monthly average maximum section passenger flow in summer is only 5,476 records and lower than other seasons, which indicates the daily travel demand of residents influenced by the weather factor excluded by basic demands.

The vehicle and passenger flow can be described as the first 7% of vehicle contributes 11% of passenger flow. The first 10% of vehicle contributes 15% of passenger flow. The first 20% of vehicle contributes 30% of passenger flow. It is known that full-load ratio of some vehicles is quite high during peak hour, and urban transit department should make some plans to ease the pressure of passenger flow such as increasing more vehicles.

![Figure 2. Monthly passenger flow and Passenger flow of bus line.](image-url)
5.2 Analysis of the Weather Impact on Passenger Flow

Date data in comfortable and non-comfortable weather is selected combined with historical weather data in Guangzhou. The statistics of average hourly passenger flow in different weather is shown in Figure 4. It is known that the average hourly passenger flow in comfortable weather is a bit higher than that in non-comfortable weather, but the gap is not more than 200 records within a day. The possible reason is the duration time of non-comfortable weather will not be too long in a day, passengers can travel in other comfortable periods to avoid the non-comfortable periods, which is also in accordance with intention of passengers.

5.3 Analysis of Types and Areas of the Passenger

The percentage of passenger type could be counted through the card type field of smart card. The passenger flow group are mainly made up of these three groups of passengers. The percentage of ordinary passenger is 77.3%. The older passenger is 15.7%. The student passenger is 5.8% (check to see Figure 5(a)). The requirements of older card that is created by Guangdong Lingnantong incorporated company is over 65 years old residents or living more than half a year from foreign residents. The population data is published by Guangzhou Civil Affairs Bureau in 2014[9]. The number of more than 60 years old residents has over 1.33 million. The number of 65-75 years old residents accounts for 37% of the older resident population. The number of students in primary and secondary school has surpassed 2 million. The older passenger with travel demand is estimated that 65-75 years old residents could buy food or do some other activities. The older passengers’ flow is three times more than the students’ passenger flow, but the number of older residents with travel demand is less than the number of the students. It is known to us all that the travel frequency of older residents is much higher than the students. One reason is that the older residents need buy vegetables and other necessities of life by bus. The other reason is that their travel distance may not be far but the frequency is a little higher.

Statistics of regional distribution of passenger flow could be analyzed by the city of card creation in the smart card records. According to the statistics related to the city were the card was bought, 92.8% of the passengers of the bus line are from Guangzhou, 6.2% are from Foshan, and 1% are from other cities. It is clear that the main passenger flows of Guangzhou bus line are local residents (see Figure 5(b)).
5.4 Cluster Analysis of Passengers Travel Frequency

$K$-means algorithm in the literature [5] is used on passenger travel frequency clustering, the clusters $K$ is set to 5, $K=1,2,3,4,5$. The travel frequency level is set as very high, high, medium, low, very low in these five levels. Distance measurement uses Euclidean. Centroid method uses random initialization, and its convergence condition is 0.1. The maximum number of iterations is 100 (see Table 6). The percentage of total for “low” and “very low” level in ordinary card, student card and older card is close or more than 80%. The percentage of total for “medium” level and above is not over 5%. It can be seen that most of the passengers will not travel a lot by bus. The public transit department could give the corresponding service so as to increase their travel frequency through the investigation of the passenger travelling demand.

### Table 1. Result analysis of passenger clustering.

| Card Type   | Analysis Index | Value            |
|-------------|----------------|------------------|
|             | Travel Frequency | Very High, High, Medium, Low, Very Low |
| Ordinary Card | Cluster Center Distance (records) | 165.07, 78.66, 33.32, 10.19, 1.74 |
|             | Number of Cards | 3,457, 11,989, 39,695, 181,694, 1,268,916 |
|             | Percentage of total | 0.23%, 0.80%, 2.64%, 12.07%, 84.27% |
| Student Card | Cluster Center Distance (records) | 141.29, 62.39, 25.47, 8.12, 1.59 |
|             | Number of Cards | 270, 1,200, 3,979, 17,073, 100,380 |
|             | Percentage of total | 0.22%, 0.98%, 3.24%, 13.89%, 81.67% |
| Older Card  | Cluster Center Distance (records) | 184.34, 77.20, 29.96, 9.76, 1.98 |
|             | Number of Cards | 424, 2,018, 9,558, 44,974, 223,093 |
|             | Percentage of total | 0.15%, 0.72%, 3.41%, 16.06%, 79.66% |

6. Conclusion

This paper makes some research on the big data of the passenger flow analysis in Guangzhou based on smart card, which is analyzed by cloud computing technology. $K$-means algorithm is used to cluster the passenger travel frequency level, which have a good application effect on urban transit department for providing service for attracting passengers’ flow. The data source is the database system of bus company and a little external data. How to get the external data source more diversified to analyze the passenger flow and more diversified evaluation results from an operational perspective will be future research directions.

References

[1] D.Y Yang., Big data: perception, cognition, and insight into the urban traffic system. Communication & Shipping, 2(2015), pp:4-7.
[2] F. Chen., J.F. Liu., Analysis of passenger flow characteristics based on IC card data -- a case study.
of Beijing City Urban Transport of China, (2016), pp: 51-59.

[3] X. Dai., Approach on the information analysis of urban public traffic based on the data of bus intelligent card. Southea University, (2006).

[4] G.H. Shu., The research on the rules of public transport passenger flow based on data mining. Beijing Jiaotong University, (2003).

[5] X. Ma., Y.J. Wu., Y. Wang, et al. Mining smart card data for transit riders’ travel patterns. Transportation Research Part C: Emerging Technologies, 36 (2013), pp 1-12.

[6] C. Morency., M. Trepanier., B. Agard., Measuring transit use variability with smart-card data. Transport Policy, 14 (2007), pp: 193-203.

[7] R.M. Xiao., B. Li., Y.S. Chen., Trend analysis of expressway transportation based on big data. Journal of Traffic and Transportation Engineering, 15 (2015), pp: 85-90.

[8] Aliyun Computing Corporation. Big Data Computing Service. https://www.aliyun.com/product/odps/

[9] Guangzhou Municipal Bureau of Statistics. Guangzhou statistical information handbook. http://data.gzstats.gov.cn/gzStat1/chaxun/ndsj.jsp.