Journal of
Information Technology & Software Engineering

Maktoubian, J Inform Tech Softw Eng 2017, 7:4
DOI: 10.4172/2165-7866.1000211

Research Article Open Access

Analyzing Large-Scale Smart Card Data to Investigate Public Transport Travel Behaviour Using Big Data Analytics

Jamal Maktoubian[1]*, Mohebollah Noori[2] Mehran Ghasempour-Mouziraji[3], Mahta Amini[4]

[1]International School of Information Management(ISIM), University of Mysore, Mysore, India.
[2]Zarghan Branch, Islamic Azad University, Zarghan, Iran.
[3]Department of Engineering, Islamic Azad University of Sari, Sari, Iran.
[4]Department of Computer Science, Shahid Beheshti University, Tehran, Iran.

Abstract

In urban public transport, Smart card data have been used more and more in order to collect fare automatically. They allowed passengers to access almost all type of public transportation system modes (bus, train, tram, funiculars, LRT, metro, and ferryboats) with a single card that is valid for the complete journey. Although Smart card major concentration is in revenue collection, they also generate massive amounts of passive data from the technological devices installed to control the operation of them. Generated data could be beneficial to transit planners which could rise the better understanding of passengers' behavioral patterns for short and long term service planning. However, one of the major challenges is the fact that traditional infrastructures and methods are inefficient when processing or analyzing a large volume of data. Thus, as an alternative, big data technology could be employed to enhance collecting, storing, processing, and analyzing the data. Moreover, the main motivation would be cost-efficiency of this methodology as the cost of processing and analyzing large-scale data is huge. This experience demonstrates that a combination of planning knowledge, big data, and data mining tool allows to produce travel behaviors indicators, public transport policies, operational performance, and fare policies.

Keywords: Big data analytics; Public transportation; Smart card data; Group travel behavior

Introduction

As the rapid advancement in information and communication technologies, large quantities and variety of social, economic, and environment-related data have been generated. Specially, exponential growth of location based services (LBS) followed by The popularity of location-acquisition technologies such as Social Networking Services (SNS) and Global Positioning System (GPS), has caused the generation of large-scale urban data. This progress has provided massive opportunities to better understand human mobility in urban areas in different dimensions. Currently, other equipment including Smart Card (SD) also help to promote intelligent public transportation (ITS) in different ways. The smart card is a small, tamperproof computer which turn into an essential component in most advanced public transport system around the world. Utilizing such cards has benefits for passengers as well as public transport operators or authorities. The cards are generally rising facilities for travelers, operators value in specific in order to decrease handling fees. Moreover, Smart cards make it much easier and More Productive to merge the fare systems of different operators and to distribute the revenues. The importance of smart card has been increasingly recognized as a rich data source to better understand demand patterns of passengers and could be interesting area for further investigation. However, by growth in the number of travelers and progress in developing public transportation, the volume of data has been dramatically enlarged than before. Because cards gather day-to-day travel behavior continuously, the size of data might become so large that after few days it is difficult to handle.

Smart card data can in this perspective be viewed as one sort of 'Big Data (BD)'. Using big data technology, we are able to consider nearly the whole system population data in analyzing passengers behavior. While in traditional data analysis, data sampling methodology require in order to select small size of data from the entire. Statistical methods such as factor analysis and/or clustering analysis are often adopted to understand the sample characteristics, but the procedure is far more difficult considering the data size. Generally, in order to apply big-data analysis in every organization the following steps require to be consider.

Data preprocessing

Data should be preprocessed through a series of specific techniques such as data cleaning, data integration, data transformation, and data reduction.

Storing data

There are various types of databases that could be employed to store data including Hadoop Distributed File System(HDFS), Hbase, Apache Cassandra, Redis, and to name but a few.

Big data processing frameworks

Big data processing could be divided into three major sectors including batch processing, real-time processing, and hybrid processing.

Big data analytics

There are some cutting-edge analytic procedures such as text mining, machine learning, predictive analytics, data mining, and natural language processing (NLP), that businesses and organization can analyze their data to obtain new insights into the problems and make faster decisions.

The paper is structured as follows. The next section provides a literature review of previous research on travel behaviour evaluations using public transport smart card data. Following is a description of the SCD-based identification method. The paper concludes with research methodology which demonstrate the different steps of research project.

***Corresponding author:** Jamal Maktoubian, International School of Information Management (ISIM), University of Mysore, Mysore, India, Tel: +989217143226; E-mail: jamal.maktoubian@gmail.com

Received October 11, 2017; **Accepted** October 24, 2017; **Published** October 31, 2017t

Citation: Maktoubian J, Noori M, Mouziraji MG, Amini M (2017) Analyzing Large-Scale Smart Card Data to Investigate Public Transport Travel Behaviour Using Big Data Analytics. J Inform Tech Softw Eng 7: 211. doi: 10.4172/2165-7866.1000211

Literature Review

This section provides a brief literature review on the uses of smart card data to study travel behavior. In 1968, a plastic card including a microchip came to Dethloff and Grotrupp mind [1]. Since 1990s, with the advent of the Internet technology and rise of mobile communication technologies, the design and use of smart card has been revolutionized [2]. Provide a literature review of the uses, which they divide into three groups: operational, tactical and strategic-level applications. Operational-level studies use smart card data to measure various transit supply-and-demand and performance indicators; tactical level studies commonly focus on service adjustments; and strategic-level studies typically relate to long-term network planning, demand forecasting and travel behavior. In this research proposal, a strategic-level analysis of travel behaviours indicators, public transport policies, operational performance, and fare policies was conducted. Agard et al. [3] presented that analyzing smart card could aim to better understanding of user behavior, since every single passenger could be followed during his/her journey.

Recent years, many researchers have started to analyze travel behavior using these intensive data. For instance, Gonzalez et al. studied people mobility patterns utilizing the trajectory of 100,000 anonymized cell phone clients in 2008 [4]. Jiang analyzed taxi commuters travel characteristics and identified the impact on urban structure on those properties utilizing GPS-recorded taxi trajectories in Sweden. Li et al. [5] investigated the relationships between tweet and photo densities and the characteristics of local people in California using geo-referenced tweets and photos aggregated from Flickr and Twitter.

Bryan and Blythe [6] discussed the possibility to analyze cardholder's behavior by using SCD. Park et al. [7] proved that the SCD has the potential as a basis for describing the characteristics of public transit users. Kang et al. [8] utilized a real-time Oyster card database of individual passenger journey in the subway of London to understand the polycentric urban London structure. Zhong et al. [9] dynamically identified the spatial structure of urban hubs and the borders using Singapore's SCD. Relevant literature review for SCD application in the public transit context can be found in Pelletier et al. [10].

Agard et al. [3] presented a typical transport planning/data mining approach for travel behavior analysis. Different measures regarding the variability of travel behaviors of transit users were proposed. Chu and Chapleau [11] indicated the use of advance statistics, sophisticated GIS analysis, visualization, and machine learning and data mining to show travel behavior of passengers. Mohamed et al. [12] showed a way to deal with analyzing the temporal behavior of the travelers in a public transportation system to obtain relevant clusters. Four weeks of journeys by both bus and subway from the urban area of Rennes in France was gathered and results were applied on a dataset. Tao et al. [13] demonstrated a multistage ways in order to render more insightful spatial-temporal patterns of urban public transport (UPT) passenger's behavior. Zhang et al. [14] proposed and implemented a novel architecture called mPat to investigate people movement utilizing various data source, including SCD feeds from 24 thousands vehicle, 16 million smart cards and 10 million cellphones. Yen et al. [15] explored the effects of the two principal features of the fare policies applied in SEQ, using smart card transaction records.

A SCD-Based Identification Method

Perhaps group walking behavior is the most well understood type of GTB. Previously, most of existing literature analyzed walking behavior by treating pedestrians as isolated individuals, each having an own desired speed and direction of motion [16]. The conceptual model can provide a concise and fundamental framework to understand Group travel behavior (GTB), however, it's unlikely to identify GTB directly based on this model. There are two main difficulties. One is the difficulty of calculating group distance, which is dynamically changed and determined by several factors that hard to be quantitatively measured. The other is that technologies nowadays cannot record interpersonal distance between individuals at any time during a continuous time duration. In addition, it's also not permitted due to privacy protection.

Most of SCD collected from the public transit smart card system in cities can tell us some basic attributes of public transit trips. These attributes usually contain the departure and arrival time, the departure and arrival stops, the line numbers of bus or subway, and card ID etc. In addition, group travelers move together and keep a very close distance at most time of a trip. What's more, implied by previous psychological studies [17,18] and our daily experience, group travelers also have a preference to swipe their cards continuously, while other strangers usually try to avoid to swipe cards during group travelers. Based on the data and facts, an identification method will be developed to identify GTB in the public transit context. The basic idea is that: if people get on the bus (or subway entrance) at a very close time at the same stop (or station), stay in the same bus (or subway) line during the trip, and get off the bus at a very close time at the same stop, they will be identified as group travelers. Staying at the same departure and arrival stops and the same transport line means they are close (or visible for each other) spatially. Based on spatial closeness, getting on or off at a very close time means they are close enough to be called as group travelers. Time interval between their swiping (called as interpersonal (time) distance here, while interpersonal distance mentioned previously in this proposal is actually interpersonal (spatial) distance) can be treated as the indicator of interpersonal (spatial) distance. When the time interval, namely interpersonal (time) distance, is small enough to reflect a group (time) distance in the swiping card situation, it can be used to identify GTB. Although it's workable for us to know whether people stay within a group (spatial) distance in the station entrance or bus stop by checking their swiping card behavior, it's impossible for us to know their interpersonal (spatial) distance in the bus or subway at any time. But group (time) distance at the starting and end time of the trip can help us "imagine" whether they stay close enough or not when they are in the transport lines.

Research Methodology

As it can be seen from the Figure 1, smart card data require to be collected from transportation authorities, and then processed, analyzed and finally taken into account for appropriate action. Generally, the development of this paper would be split up into three steps:

Data collection

Smart card data (SCD) which are produced by smart card systems are one type of urban big data since the number of travelers who are fond of employing smart cards during their journeys have been risen significantly. The SCD track the detailed onboard transactions of each passenger, and can create a complete and real-time journey diary for all vehicles travels. Fare collection systems are more and more applied in public transit systems; progressively SCD have been generated and been accessible for research reason or public use. Hardware requirements for developing a big data platform should be provided and then data require to be stored in Hadoop Distributed file system (HDFS) or any other environment (HBase, Apache Cassandra, etc.) which is capable to handle terabytes data. However, before processing data, they should be go through different steps which is called "Data Cleaning". In order to cleansed data, various stages require to be considered, which include filling in missing data, smoothing the noisy data, or resolving the inconsistencies in the data. Raw data should to be preprocessed in four

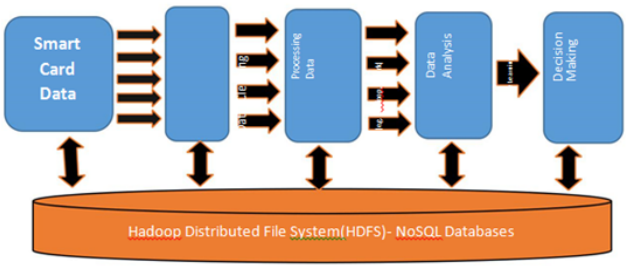

Figure 1: General architecture.

major steps:

Data integration: Data with diverse and complex representations are put together and conflicts within the data are resolved.

Data transformation: Data is converted, normalized, aggregated and generalized.

Data reduction: This stage aims to present a decreased data representation in a data warehouse.

Data discretization: Includes the decrease of various values of a continuous quality by splitting the range of attribute intervals.

Data processing

Cleaned data need to be processed by relevant big data processing technology such as Apache Hadoop, Spark, Storm, and to name but a few. Generally, big data processing could be arranged into two types: batch processing, and real-time stream processing. Based on the type of data and their needs, scientists and developers decide what processing methodologies would be fit for their system. Since the analyzing of passengers 'behavior who are using public transportation smart card needs stored data, batch processing strategies will be employ as a processing technique. Apache Hadoop [19] is an open-source framework which has two main parts, the first division is distributed storage (Hadoop distributed file system (HDFS)) and the second one is processing technique (MapReduce programming) of large datasets across clusters of nodes(computers). Apache Spark is another big data processing platform which [20] is fast, in-memory data processing engine, and an open-source, with built-in modules for streaming.

Data analysis

Machine learning, data mining algorithm and analytic methods could help us to explore passenger trip purpose, identify transit use cycle and travel patterns among card segments and to measure travel patterns for different types of passengers. Big data platform such as Apache spark could make this operation much faster and easier. Spark MLlib, Spark's Machine Learning library, contains different machine learning algorithms and techniques designed to scale out on a Filtering, dimensionality reduction, Clustering, Classification and other methods. At a basic level, statistical methods for the analysis of passenger travel patterns include frequency analysis, ANOVA and related spatial and temporal correlations among journeys. And in advance, using clustering methods, we could look at temporal and spatial travel patterns, usually by origin and destination and by time of the day. K-means clustering would be ideal to identify the typical spatial and temporal travel patterns and to identify "anomalous" behaviour that does not easily fit existing clusters. A Naïve Bayes classifier could be utilized to classify passenger trips based on the day of week, time of day and frequency of travel. An extension of this model to predict passenger boarding sites. Generally, using big data technologies in analytics would aim to increase the accuracy of results and find other hidden pattern which required more data.

Conclusion

Since the amount, variety and complexity of data have been increasing dramatically, traditional methods are failed to collect, process and analyze data. The first objective of this paper is to present that emerging Big data technology could be employed to measure various types of travel behavior. In this paper, the major steps from collecting data to data analysis are discussed.

References

1. Shelfer KM, Procaccino JD (2002) Smart card evolution. Communications of the ACM 45: 83-88.
2. Blythe PT (2004) Improving public transport ticketing through smart cards. Proceedings of the Institution of Civil Engineers-Municipal Engineer 157: 47-54.
3. Agard B, Morency C, Trepanier M (2006) Mining public transport user behaviour from smart card data. CIRRELT 39: 399-404.
4. Gonzalez MC, Hidalgo CA, Barabasi AL (2008) Understanding individual human mobility patterns. Nature 453: 479-482.
5. Li L, Goodchild MF, Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cartogr Geogr Inf Sci 40: 61-77.
6. Bryan H, Blythe P (2007) Understanding behaviour through smartcard data analysis. Proceedings of the Institution of Civil Engineers-Transport 160: 173-177.
7. Park J, Kim DJ, Lim Y (2008) Use of smart card data to define public transit use in Seoul, South Korea. Transp Res Rec 2063: 3-9.
8. Roth C, Kang SM, Batty M, Barthelemy M (2011) Structure of urban movements: polycentric activity and entangled hierarchical flows. PloS one e15923.
9. Zhong C, Arisona SM, Huang X, Batty M, Schmitt G (2014) Detecting the dynamics of urban structure through spatial network analysis. Int J Geogr Inf Sci 28: 2178-2199.
10. Pelletier MP, Trépanier M, Morency C (2011) Smart card data use in public transit: A literature review. Transportation Research Part C: Emerging Technologies 19: 557-568.
11. Chu K, Chapleau R (2010) Augmenting transit trip characterization and travel behavior comprehension: Multi day location-stamped smart card transactions. Transp Res Rec 2183: 29-40.
12. Mohamed K, Come E, Baro J, Oukhellou L (2014) Understanding passenger patterns in public transit through smart card and socioeconomic data. UrbComp.
13. Tao S, Corcoran J, Mateo-Babiano I, Rohde D (2014) Exploring bus rapid transit passenger travel behaviour using big data. Appl Geogr 53: 90-104.
14. Zhang D (2014) Exploring human mobility with multi-source data at extremely large metropolitan scales. Proceedings on Mobile Computing and Networking 201-212.
15. Yen BT, Tseng WC, Chiou YC, Lan LW, Mulley C, et al. (2015) Effects of Two fare policies on public transport travel behaviour: Evidence from South East Queensland, Australia. J EASTS 11: 425-443.
16. Moussaid M, Perozo N, Garnier S, Helbing D, Theraulaz G (2010) The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PloS one 5: e10047.
17. Cheyne JA, Efran MG (1972) The effect of spatial and interpersonal variables on the invasion of group controlled territories. Sociometry 35: 477-489.
18. Hadoop A (2011) Apache hadoop. Cloudera.
19. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, et al. (2016) Apache Spark: A unified engine for big data processing. Communications of the ACM 59: 56-65.
20. Polzer U (2011) Nonverbal behavior in public space as a function of density and group size.