ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 15, March 2017

Travel Pattern Mining In Smart Card Data

Elizabeth Scaria
Computer Science Department
M A College of Engineering and Technology
India
elizabethscaria24@gmail.com

Abstract— Different classes of transit users are targeted by different transit operators for targeted surveys and various operational and strategic planning improvements. This is enabled by the transit passenger segmentation. Many passenger surveys have been generally done in all the existing market segmentation methods. But these methods are having some limitations. The smart card (SC) data from an automated fare collection system facilitate the understanding of the travel pattern of various transit passengers and can be used to segment them into identifiable groups of similar behaviors and needs. The paper proposes a methodology for passenger segmentation in smart card data. After reconstructing the travel itineraries from smart card transactions, this paper adopts the density-based spatial clustering of application with noise (DBSCAN) algorithm to mine the travel pattern of each smart card user. A priori market segmentation approach then segments transit passengers into three identifiable types. An Optimized Parameter DBSCAN based methodology is proposed. The methodology proposed in this paper helps the transit operators to learn the needs of passengers and furnishes them with oriented services and information.

Keywords— ***Automated fare collection (AFC) system; market segmentation; clustering: smart card (SC); transit passenger; DBSCAN; Spatial pattern; Temporal pattern.***

I. INTRODUCTION

The transit authorities should have a better understanding of their passengers, in order to satisfy the customer requirements and preferences effectively. Many recent works on this been done and most of them have defined classes of customers but not market segments. For example, the transit operators in South East Queensland (SEQ), Australia, classify passengers according to their age and occupation (such as adult, senior, child, pension, secondary school student and student). Although this method of classification is useful in fare collection, how differently these groups respond to alternative services and new policies is unknown. And also to know whether new policies benefit them is difficult. The transit operators have only limited knowledge about their customers due to population, obscurity of passengers and ambiguity of their behaviour. Existing service improvements are limited to the effect on generic transit customers, neglecting the differences between the segments of passengers with varied requirements and behaviours. The most of studies in public transport wholly focus on improving a vehicle's performance, such as the schedule adherence or travel time, without a deep understanding of passenger types and behaviours, not withstanding the fact that different segments of passengers in a transit system would behave differently. For instance, an irregular transit passenger would be more concerned with the service coverage, i.e., if s/he would be able to travel by public transport to the desired destination, whereas a commuter transit rider user would be more concerned with the on-time performance and the easiness of transfers. Earlier studies on passenger travel patterns and passenger segmentation wholly focussed on the use of transit user surveys. These surveys are generally are limited in sample size, expensive to conduct and are only valid within the study period. Transit agencies are at a crucial transition in data collection technology from manual data collection towards Automated Data Collection Systems (ADCSs). Automated Data Collection Systems such as automatic vehicle location, automatic passenger counter, automatic vehicle identification, and, particularly, smart card (SC) automated fare collection (AFC) systems with low marginal cost and large sample sizes have replaced manual data collection systems with a low capital cost but a high marginal cost, small sample sizes, and sometimes unreliable accuracy. ADCSs is becoming widely popular for collection and analysis. The advancement of these modern technologies provides a tremendous opportunity to facilitate agencies to enhance the service quality and analyze the present condition of transit quality of service. Public transport agencies can take advantage of massive data to ensure that the strategies are effective to augment the transit experience and provides a demand-responsive transit system. And thus would earn an utmost advantage to attract customers. This paper accrues the transit passenger characterization by passenger segmentation using the dynamic Smart Card data. The aim of segmentation is to classify passengers of similar travel pattern, i.e., with the same level of transit journeys at regular times and places. The market segmentation of transit passengers brings various aids to transit authorities to provide better to their customers. Travel pattern mining helps in understanding the evolution of passenger demands, providing Incentives and personalized service to passengers of regular usage for encouraging them to use public transport. The analysis of the travel pattern also aids operational strategies such as origin–destination (OD) demand management and transfer coordination by monitoring and inferring passenger movements through their travel customs. The paper proposes a systematic methodology to mine the travel pattern and segment transit passengers using SC data. And efficient DBSCAN method is used for the same.

After the related studies in Section II, reconstruction of completed journey of SC users from individual SC transactions is discussed in Section III B. Each journey is defined as the travel from an origin to a destination, which might include one or several transactions. Each transaction includes both the boarding and alighting times and stop IDs of a transit journey between a touch on and a touch off to the ticketing device In Section III-C, a density-based clustering algorithm and optimized parameter DBSCAN algorithm are adopted to mine the travel pattern from each SC user's historical itineraries and to identify the spatial OD that the cardholder usually travels as "regular OD" and the time of regular travels as habitual time. In Section III-D SC users are segmented into different classes using the mined travel pattern by the priorimarket segmentation method. The segmentation results are analyzed and comparison of DBSCAN and Optimized DBSCAN approach is done in Section IV, the conclusion summarizes the paper.

II. RELATED STUDIES

The intelligent transportation system AFC system uses Smart Cards which can captures massive volume of travel data and assists economical, scalable and efficient method in exploring travel behaviors of transit passengers. Many recent studies on travel pattern mining has been published using SC data, where the authors have connected individual SC boarding/alighting records to reconstruct user itineraries [5]–[7]. Existing works have been looking at general passengers, individual passengers [5] and group of passengers [9]. Even these methods provide an overview of travel pattern of a general user, it is not efficient in acquiring the individuality of travel behavior. The typologies of passengers and trips are predefined, where the similarity of passengers between the same class and the difference between classes may not be reflected. Another approach is the development of pattern discretization. Spatial travel pattern analysis often breaks down to stop-to-stop repeated trips, [9]-[12]. A temporal pattern is defined if the passenger repeatedly made multiple trips within a time period. It is arduous to discretize the temporal pattern for individual passengers because different people are having different habitual behaviors. There are researches on segmentation of transit passengers based on travel behaviors for fare elasticity. Some studies used three approaches: a) physical segmentation based on information like demography, geography b) product usage segmentation based on frequency of use c) physiological segmentation based on the characteristic of individual passengers. In [1] approach similar to this paper has been done but it has used DBSCAN (Density Based Spatial Clustering of Application with Noise). Two parameters Eps and MinPts are required to be inputted manually in DBSCAN algorithm, and this tedious intervention leads to the situation that the clustering precision depends largely on user's entry. This paper uses a method to determine the two parameters, which can avoid the manual intervention, and even realize the clustering automatically. Experimental results show that the method can determine the two parameters more reasonably. Furthermore, it can get clustering results more accurately.

III. PROPOSED METHODOLOGY

In this section the data set used, as well as the methods for the reconstruction of travel itineraries travel pattern analysis using DBSCAN and Optimized parameter DBSCAN and passenger segmentation is introduced.

A. *Data Set*

The Smart Card data used in the paper is from Translink (the transit authority of SEQ, Australia). Each transaction is having the following fields:

a) CardID: The unique SC ID
b) T_on: The timestamp for touch on.
c) T_off: The time stamp for touch off.
d) S_on: The station ID at touch on.
e) S_off: The station ID at touch off.
f) ValidIndicator: A binary indicator for differentiating a valid or invalid transaction. A valid transaction is the combination of a touch on and a touch off from the same transit line within a 2-hour limit[15] . Any cases other than that includes no touch off, touch off at a different line, etc., are indicated as invalid transactions.
g) RouteUsed: The transit line that the passenger has used.
h) Direction: The direction of travel (inbound/outbound).
i) Fare: The fare paid for the transaction in Australian dollars.

B. *Reconstruction Of Travel Itenaries*

The Reconstruction of Travel trips from the individual transaction is the first step in travel pattern mining. The Reconstruction algorithm uses ReconstructingIndicator to identify the ongoing/new trip status and on a TripID to identify the completed trips. A connected transactions are decided using a fixed threshold of 1 hour. The first boarding stop of a completed trip is defined as the "origin stop" and the last alighting stop of a completed trip is defined the "destination stop". The transferring time is defined as the interval between the alighting time of a transaction and the boarding time of the next transaction of the same. [6] discussed about a method, In vehicular ad hoc networks (VANETs), because of the nonexistence of end-to-end connections, it is essential that nodes take advantage of connection opportunities to forward messages to make end-to-end messaging possible. Thus, it is crucial to make sure that nodes have incentives to forward messages for others, despite the fact that the routing protocols in VANETs are different from traditional end-to-end routing protocols. In this paper, stimulation of message forwarding in VANETs is concerned. This approach is based on coalitional game theory, particularly, an incentive scheme for VANETs is proposed and with this scheme, following the routing protocol is in the best interest of each node. In addition, a lightweight approach is proposed for taking the limited storage space of each node into consideration.

Reconstruction of Travel Itineraries process involves the following steps:

S1: A binary ReconstructingIndicator is defined and initialised as 0.

S2: The ValidIndicator is checked. If the indicator is equal to 0 the transaction is invalid and the corresponding trips will be discarded.

S3: If the ReconstructingIndicatoris 0, a variable OriginLocation is defined and set as equal to the current T_on. A new unique TripID is assigned and the ReconstructingIndicator is changed to 1, save the current transaction, and move to the next transaction. If the ReconstructingIndicator is 1 and the time gap between the current T_on and the last T_off is less than 1 hour, we move to S4. If the time gap is more than 1 hour, the transaction with the previous TripID is connected into a completed trip. A new TripID and a new OriginLocation are assigned. The ReconstructingIndicator is set as 1.

S4: If the current S_off is different to the Origin- Location, the transaction is connected to the trip as a continuation journey. If it is also the last transaction of the day, the trip reconstruction process for the study passenger is finished; otherwise, we move to the next transaction.

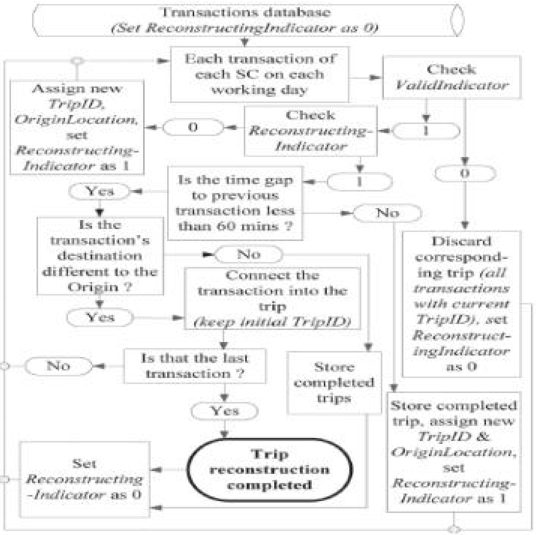
Transactions database
(Set ReconstructingIndicator as 0)
Each transaction of each SC on each working day
Check ValidIndicator
Assign new TripID, OriginLocation, set Reconstructing-Indicator as 1
0
Check Reconstructing-Indicator
1
0
Yes
Is the time gap to previous transaction less than 60 mins ?
1
No
Is the transaction's destination different to the Origin ?
No
Connect the transaction into the trip (keep initial TripID)
Discard correspond-ing trip (all transactions with current TripID), set Reconstruct-ingIndicator as 0
Yes
No
Is that the last transaction ?
Store completed trips
Yes
Store completed trip, assign new TripID & OriginLocation, set Reconstructing-Indicator as 1
Set Reconstructing -Indicator as 0
Trip reconstruction completed

Figure 1 Trip Reconstruction Flow chart

C. Mining Spatial and Temporal Pattern from Travel Itineraries

This section introduces the method to mine the spatial and temporal travel pattern from the historical trip database. The spatial OD stops are represented as geographical coordinates whereas the temporal boarding and alighting times are represented as timestamps. Density-based clustering algorithm is adopted because Density-based algorithms can identify clusters of high density and noise of low density thus regular patterns can be identified and can be differentiated from anomaly pattern, Density-based algorithms can identify a cluster of any shape and size ,and it do not require the predetermination of initial cores or the number of clusters. Density- based clustering algorithms produces high density clusters for spatial and temporal patterns.

There are many density-based clustering algorithms such as the density-based spatial clustering of application with noise (DBSCAN) [16] and more complex methods such as ordering points to identify the clustering structure (OPTICS) [17] and density-based clustering (DENCLUE) [18] can be found in literature. DBSCAN is then chosen as the algorithm to use in this paper because of its high computing performance to handle a large data set with over a million SC users. Two parameters Eps and MinPts are required to be inputted manually in DBSCAN algorithm, and this tedious intervention leads to the situation that the clustering precision depends largely on user's entry so this paper uses optimised parameter DBSCAN which adaptively computes these parameters thus eliminating the limitation.

1) DBSCAN Algorithm:

The DBSCAN algorithm defines clusters as dense regions, which are separated by regions of a lower point density. The algorithm has two global parameters: the maximum density reach distance ε and the minimum number of points MinPts and it will output C number of clusters. A set of points in some space is grouped together into points that are closely packed together, points with many nearby neighbors, marking as outliers points that lie alone in low-density regions whose nearest neighbors are too far away.

2) Optimised parameter DBSCAN Algorithm

In DBSCAN clustering algorithm, the two parameters of Eps and MinPts are needed to be inputted manually in advance. This leads to the situation that the clustering accuracy depends on user's selection of parameters. And furthermore, its time complexity is close to O(n^2) in the worst case . In view of these, this paper a new scheme [2] to decide the values of Eps and MinPts through analyzing the distribution properties investigated objects. In such way, the whole clustering process can be fully automated. The main idea of this algorithm is the values of parameters Eps and MinPts are ascertained based on the statistical properties of the data set.

A distance distribution matrix DISTn×n needs to be calculated in advance.

$$\mathrm{DISTn} \times \mathrm{n} = \{\mathrm{dist}(\mathrm{i}, \mathrm{j}) \mid 1 \leq \mathrm{i} \leq \mathrm{n}, 1 \leq ?\, \mathrm{j} \leq \mathrm{n}\}$$

n means the number of objects in the data set D. DISTn×n is a real symmetric matrix with n rows and n columns, in which each element denotes the distance between objects i and j in D.

We use the maximum likelihood estimation in mathematics to estimate the value of parameter Eps. That is to say, Eps can be obtained by means of the geometrical mean of the value of DISTn×i.

$$Eps = \frac{1}{n} \sum_{i=1}^{n} X_i$$

After Eps is determined, the number of data objects in Eps neighborhood of every point in dataset is calculated one by one. And then mathematic expectation of all these data objects is calculated, which is the value of MinPts

$$Minpts = \frac{1}{n} \sum_{i=1}^{n} P_i$$

where Pi is the number of points in Eps neighbourhood of point i.

The algorithm is separately applied for mining the spatial and temporal patterns, in which the regular ODs are derived by a two-level DBSCAN application. The separate application of DBSCAN increases the robustness of the overall clustering algorithm, and the outcomes of each level are used for the later passenger segmentation process.

D. Transit Passenger Segmentation

The identifiable passenger classes are selected from the SC user population based on the proportion of regular OD/habitual time trips in the total transit usage. The passenger travel characteristics, i.e., spatial and temporal travel patterns are used to define the type of passengers. During the passenger segmentation process each SC user itinerary is revisited. Passengers travelling during the study period for a certain number of journeys follows a regular OD, a habitual time pattern, or not following any pattern. Three segments of passengers can be identified based on the following heuristic rules. Only passengers with no recognizable pattern are segmented into the irregular passenger type. The other passengers could be grouped into two identifiable classes.

Rule 1: If no temporal or spatial travel pattern is identified, the passenger is classified as an irregular passenger.
Rule 2: If more than 50% of the journeys were made within habitual times and between regular ODs, the SC user is classified as a transit commuter.
Rule 3: The remaining passengers are segmented into regular OD passengers if the proportion of the regular OD journeys is more than the habitual time journeys, and vice versa for the habitual time passengers.

IV. EXPERIMENTAL RESULTS

Using Base method nine spatial clusters and ten temporal clusters were formed whereas using proposed method four spatial clusters and twenty four temporal clusters were formed. The input parameters ε and Minpts to DBSCAN for temporal clustering where 150 and 8 respectively and for spatial clustering it is 1300 and 6. Whereas proposed method adaptively compute ε=2012, MinPts=8 for temporal clustering and ε=132, Minpts=6 for spatial clustering. The accuracy of methods were compared using calculating the F-measure (F1-score)

$$F1 - Score = 2 \times \frac{precision \times recall}{precision + recall}$$

where precision and recall can be defined as

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)}$$

$$Recall = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)}$$

Experimental result shows that the proposed method has more F1-Score than the base method. Thus it is having more clustering accuracy. The segmentation results are given in the table.

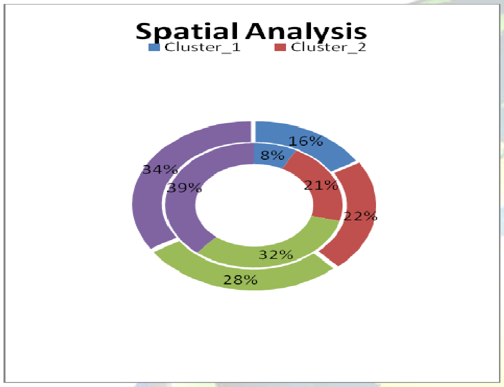

Figure 2 Comparison of F measure of spatial analysis of base and proposed method

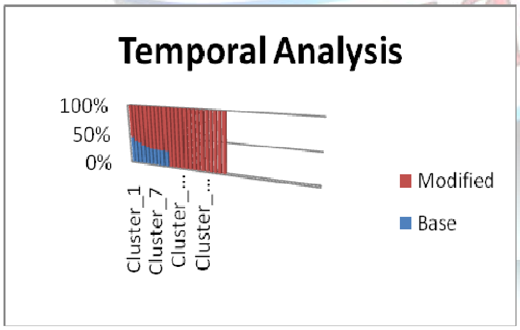

Figure 3 Comparison of F measure of temporal analysis of base and proposed method

TABLE 1 RATIO OF TRANSIT AND REGULAR COMMUTERS IN CLUSTERS

Ratio of Transit Commuters.	Ratio of Regular Commuters.
34:17	51:25
35:17	49:24
33:16	48:24
39:19	55:27
37:18	46:23
40:20	44:22
41:20	-
30:15	-

V. CONCLUSION

There are various clustering techniques available with varying attributes which is suitable for the requirement of the data being analyzed. In this paper, passenger segmentation using smart card data, optimised parameter DBSCAN is particularly chosen because it can identify clusters of high density and noise of low density without need of predefined parameters. The contribution of this study is twofold: First, a data mining approach has been proposed that is capable of identifying travel patterns for individual transit riders using a large smart card dataset. The second contribution is that the regularity levels for the data can also be successfully classified by the approach proposed here. The travel patterns and regularity levels of their customers are important information for transportation researchers seeking to understand day-to-day travel behaviour variability and facilitate activity-based travel demand model development. Individual travel patterns and pattern regularity also offer substantial benefits for transit agencies working to improve their transit service with the assistance of transit market analysis.

References

[1] Le Minh Kieu, Ashish Bhaskar, and Edward Chung, "Passenger Segmentation Using Smart Card Data," IEEE Trans. on intelligent transportation systems, vol. 16, no. 3, june 2015

[2] Hongfang Zhou, Peng Wang, Hongyan Li, "Research on Adaptive Parameters Determination in DBSCAN Algorithm", Journal of Information & Computational Science 9: 7, 2012.

[3] D.A. Hensher, "Establishing a fare elasticity regime for urban passenger transport," J. Transp. Econ. Policy, vol. 32, no. 2, pp. 221–246, 1998.

[4] Y. Shiftan, M. L. Outwater, and Y. Zhou, "Transit market research using structural equation modeling and attitudinal market segmentation," Transp. Policy, vol. 15, no. 3, pp. 186–195, May 2008.

[5] Transport Plan for Brisbane 2008–2026, Brisbane City Council, Brisbane, U.K., 2008. [5] K. K. A. Chu and R. Chapleau, "Augmenting transit trip characterization and travel behavior comprehension," Transp. Res. Rec.—J. Transp. Res. Board, vol. 2183, pp. 29–40, 2010.

[6] Christo Ananth, Kavya.S., Karthika.K., Lakshmi Priya.G., Mary Varsha Peter, Priya.M., "CGT Method of Message forwarding", International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE), Volume 1, Issue 1, August 2015,pp:10-15

[7] C. Seaborn, J. Attanucci, and N. Wilson, "Analyzing multimodal public transport journeys in London with smart card fare payment data," Transp. Res. Rec.—J. Transp. Res. Board, vol. 2121, pp. 55–62, 2009.

[8] J. M. Farzin, "Constructing an automated bus origin-destination matrix using farecard and global positioning system data in Sao Paulo, Brazil," Transp. Res. Rec.—J. Transp. Res. Board, vol. 2072, pp. 30–37, 2008.

[9] M. Utsunomiya, J. Attanucci, and N. Wilson, "Potential uses of transit smart card registration and transaction data to improve transit planning," Transp. Res. Rec.—J. Transp. Res. Board, vol. 1971, pp. 119–126, 2006.

[10] K. K. A. Chu, R. Chapleau, and M. Trepanier, "Driver-assisted bus interview," Transp. Res. Rec.—J. Transp. Res. Board, vol. 2105, pp. 1–10, 2009.

[11] S. Lee and M. Hickman, "Trip purpose inference using automated fare collection data," Public Transp., vol. 6, no. 1/2, pp. 1–20, Apr. 2014.

[12] C. Morency, M. Trepanier, and B. Agard, "Measuring transit use variability with smart-card data," Transp. Policy, vol. 14, no. 3, pp. 193–203, May 2007.

[13] S. Lee and M. Hickman, "Are transit trips symmetrical in time and space?" Transp. Res. Rec.—J. Transp. Res. Board, vol. 2382, pp. 173–180, Dec. 1, 2013.

[14] S. G. Lee, M. Hickman, and D. Tong, "Stop aggregation model," Transp. Res. Rec.—J. Transp. Res. Board, vol. 2276, pp. 38–47, 2012.

[15] How to Use Your Go Card on the TransLink Network: TransLink Go Card User Guide (Part 1 of 2), Translink, Vancouver, BC, Canada, 2007.

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowl. Discov. Data Mining, 1996, pp. 226–231.

[17] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," ACM SIGMOD Rec., vol. 28, no. 2, pp. 49–60, Jun. 1999.

[18] A. Hinneburg and D. A. Keim, "An efficient approach to clustering in large multimedia databases with noise," in Proc. KDD, 1998, pp. 58–65.