International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 08 | Nov-2015 www.irjet.net p-ISSN: 2395-0072

Passengers Segmentation for Metro Run using Smart Card

Ayushi Gaur*, Nidhi Tawra, Dharmveer Singh Rajpoot

Department of Computer Science and Engineering
Jaypee Institute of Information Technology, Noida, India

Abstract – ***This paper introduces an extensive analysis of passenger segmentation by using Smart card data. Smart card systems maintain large amount of transaction data which can be further utilized for the segmentation purpose. A mining methodology is used which basically pit the tour pattern of the passengers on the basis of the time that a passenger waits for another train, how much time passenger spends in the train , total transfer time etc. This analysis can be useful in synchronizing the number of trains on a most or least crowdie route, travelling pattern of passengers, travel time prediction and travel planning. This paper uses a Bayesian decision tree algorithm for mining the travel pattern of the passengers and a priori market segmentation algorithm for segmentation of smart card holders.***

Key Words: *Smart Card, Market Segmentation, Metro Systems, Passenger, Data Mining etc.*

1. INTRODUCTION

Nowadays, metro systems [1] have become one of the most advanced public services, if we compare this System with others, it has the assistance of high efficiency, large capacity, and fast speed .Metro passenger market segmentation provides travel authority to target divergent type of users for targeted survey. Nevertheless, the current market segmentation studies in the observation have been basically done using user surveys, having many restrictions. The smart card data from an electronic fare compilation system enables the understanding of the travel pattern of passengers and can be used to segment passengers into different types of similar nature and requirement of travelers. Better understanding of passengers is essential for higher transit authorities to satisfy customer needs and priority. Despite the high exposure to transit passengers, transit authorities have little knowledge about their customers due to reasons such as the anonymity of passenger's behaviors and the problem in analysis of the distributed information of massive population. Existing systems are limited to the impacts on generic customers, ignoring the differences between the types of passenger's behaviors and their needs. This paper augments the transit passenger simulation by passenger segmentation using the Smart Card data. The segmentation aims to cluster passengers of similar travel pattern, i.e., with the same type of journeys at usual times and locations. The segmentation of passengers brings out various transit authorities to gratify their customers.

1.1 OVERVIEW

The key for efficacious segmentation is to examine the need to analyze, at wide level. Progressively, the basic paradigm to mine travelers are reason for tour, total time spend, regularity of travelers and their engagement to the environment. More explicitly, for a given tour, we want to segment it to diverse traveler segmentations, with location and transfer time data both. For most advanced metro setup, only each tour's entry and exit time can be explicitly retrieved and all other middle points are unknown. By deducting the entry time from the exit time the total tour time can be obtained. Given only the tap-in and tap-out time, the problem is how to reduce the time period between two consecutive stations.

1.2 APPLICATIONS

Our survey can be used in various applications. Metro has become a dire need for travelling. It saves a lot of time, can be affordable and suits for every type of passenger. Some of its applications are described below.

- The understanding of every passenger type helps travel authorities in travelling strategic doctrinaire. Transit-on-demand services that serve passengers who require regular travel where standard routes are not possible can be developed.
- Transfer coordination may be developed for major transfer stops used by large numbers of straphanger and routine passengers.
- The number of travelers at each type before and after a policy implementation is a significant evaluation of various fares [7], marketing, and servicing strategies. For instance, more travelers and less intermittent passengers mean that more travelers become daily users of public vehicles.

1.3 WELL KNOWN METHODS

This section defines the methods which can be availed for the segmentation purpose. We will use a non parametric approach which is Bayesian decision tree approach instead of density based algorithm which is a parametric approach for mining travel pattern , after that we will use a priori approach for segmentation of travelers.

1.3.1 BAYESIAN DECISION TREE ALGORITHM

Bayesian decision theory uses Bayesian probability for deciding which attribute should become a node in the tree. It is a statistical system that in which various decisions are made based on various calculations to find out best tree suited, making use of probabilities and costs. A concept of Bayesian statistics is used to estimate the expected value. These agents are called estimators. This algorithm uses a non parametric approach so can be used for making decision on large dataset.

1.3.2 DENSITY BASED ALGORITHM

The DBSCAN algorithm can classify clusters in wide spatial data sets by confirming at the provincial density of database members, takes only one parameter as input. Additionally, the user gets a notification on which parameter assessment would be more suitable. Consequently, least information of the domain is necessary. The DBSCAN can also classify what value should be stated as noise or outliers. Despite, it is fail to manage the local density variation that present within the cluster. It does not work well in case of high dimensional data.

1.3.3 A PRIORI MARKET SEGMENTATION Analysis for Transit Passenger Segmentation[7]

Apriori algorithm is used to find frequent item set and association rule learning over transactional database. After finding frequent item set we find association rules. This is done based on support count and confidence value. Our aim is to get 100% confidence but we decide a certain threshold value for that same as for support count too. This algorithm uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. The Apriori Algorithm is an influential algorithm for frequent item sets mining for Boolean association rules. A priori algorithm uses a "bottom-up" approach, where frequent item sets are taken one item at a time and groups of candidate item sets are checked against the data.
In this paper, we propose an efficient method of passengers segmentation for metro run based on smart card.

Rest of the paper is organized as follow; Section 2 provide the valuable literature survey on the topic, Section 3, briefly describe the propose method then Section 4, shows the qualitative analysis of proposed method as compare to existing method, after that Section summarize the work in the form of conclusions.

2. LITERATURE SURVEY

This system uses large amount of smart card data holders to find out the day today behavior of travelers. This is done to know the travel patterns of the passengers. There will be some ease for the passengers after this survey because numbers of trains can be increased or decreased accordingly.

2.1 OVERVIEW

Non parametric approach algorithm (Bayesian decision tree algorithm is being used) so that we can work on any number of data set. Moreover, A priori algorithm is used to find frequent item sets to categorize the passengers according to their behavior. This algorithm is very easy to implement so we used this.

a) Mining Spatial and Temporal Pattern From Travel Itineraries

Spatial and temporal method is used for mining the data. For this purpose DBSCAN algorithm is used.
Cluster of any shape and size can be identified using this algorithm. Predetermination of initial number of clusters is not required. For high dimension and varying density data this algorithm doesn't work well.

b) Data mining for transit stop recognition

For this purpose Bayesian decision tree algorithm is used. This algorithm uses a non parametric approach nad using this algorithm we can work for any number of data even for the large datasets it works. There is no correct way to choose a prior using this algorithm. It often comes with high computational cost.

c) A Priori Market Segmentation Analysis for Transit Passenger Segmentation [7]

This algorithm is used to find out frequent item sets.
It is easy to implement, easily parallelized and uses large item sets property. It requires many database scans, transaction database is memory resident with this algorithm.

3. PROPOSED METHOD

For the mining of passengers effectively we are using a nonparametric approach because for large and dense data sets it is not always possible that we can calculate the distance function for each node i.e. Euclidean distance or

other distance function, So we are using an non parametric approach in which we didn't consider the distance function which implicitly perform the feature selection. After mining the passengers we use a segmentation approach i.e. A priori approach which is suitable for large datasets to mine passengers by using some threshold.

Algorithm Proposed: Initially the algorithm that we are propounding encompasses two basic steps:

Step 1:-Naive Bayesian decision tree approach for mining the travel pattern of the each Smart Card holder.

- Compute total entropy of the target by using formula
  $E_{total}=-\sum p*\log_2 p$
- Compute entropy of each branch and find Information Gain.
  Information Gain(branch)=E_{total} - E_{branch}
- Choose the attribute which is having maximum Information Gain as root and split again
- Apply these steps successively until we get a target output .

Step 2:-A priori market segmentation approach for segment passengers into different types i.e. Stable traveler, Daily traveler, Persistent traveler and Occasional traveler.

- Find frequent item sets on the basis of minimum support count.
- From frequent item sets identify association rules which satisfy minimum confidence criteria.

Advantages of proposed System:-

- This proposed method uses a non-parametric approach i.e. no distance function is used here so clustering or mining of travelers is more efficient.
- Suitable for large datasets and does not affect the performance of algorithm if there is a non linear relationship between different attributes.

4. EXPERIMENTAL RESULTS

This section of paper defines the result after implementing all the steps.

4.1 DESCRIPTION OF DATASETS

Data is taken from the Washington Metropolitan Area Transit Authority from metro planning blog on which data is available to download. Metro data is of periodic temperament, so any correlative analyses should be performed on a time span basis (i.e. October 2014), transaction data composed of following attributes:

- **Number Rider–** Traveler number who is travelling.
- **Service Type** – type of service on the basis of weekdays and weekends.
- **Entry Time Period –** entry time of a traveler that can be AM Peak, Midday, Evening, PM Peak.
- **Entry Station-** defines entry station from which a traveler enters.
- **Exit station -** station from where a traveler exists.
- **Year Month –** defines month and year of the traveler.
- **Travel Minute-** Total number of minutes a traveler spends on metro.

4.2 DESCRIPTION OF QUALITY MEASURES

Quality measures are the medium that tends to amplify the outcomes, perceptions and organizational structures.

1) **Transfer time**
Transfer time is the total one way or round trip time taken between two stations. Transfer time is generally less than 90 seconds but relative transfer time is high.

2) Segmentation
Spatial- temporal segments of trips can be find out using boarding and transfer time. For all expeditions in our database average percentile values are used.

3) Time sensitivity
Many travelers are time sensitive in the dataset. Among then no one travels in public transport .These passengers are capable.

4) Total Travel
Total travelling time is similar among different passenger types. Irregular travelers spend more time in travelling than the regular ones. This happens because of transfer time.

5) Modes and Routes
This tells us the mode of travelers and the route like bus, train. There are some passengers which uses bus only or train only. There are some passengers which uses bus-train both.

6) Accuracy
We can measure the performance and efficiency of our algorithm using accuracy, which can be calculated by
Accuracy = correctly identified unlabeled stops/unlabeled stops
For each traveler, we calculate the accuracy after running a test for each method. Overall accuracy is calculated by an average of decimal cross-verification.

4.3 QUALITATIVE RESULT AND ANALYSIS

The interpretation of each traveler type amplifies the traveler impersonation. The following interpretation about travelers could be attained from the passenger segmentation determination:

- Majority of Smart Card holders are Occasional travelers who does not pursue any legitimate route for traveling. These travelers do not earn much profit and cannot be beneficed by the different policies by transit Authority for the regular travelers.
- After Occasional travelers, Stable travelers are found to travel in regular basis which follows the same path daily in Peak Hours. It means that they are regular travelers who are beneficed by the different schemes.
- Daily Travelers are those travelers who travel daily from same entry and exit station but there is no fix time for their travel. These travelers basically travel in peak hours.
- Persistent Travelers are those who travel in the same time period but their entry and exit stations are different in different days. These travels to different destination so these are having more mobility than other type of passengers.

Stable Traveler	**Daily Traveler**
Persistent Traveler	**Occasional Traveler**

5. CONCLUSION

This paper has proposed an organized approach to mine the pattern of travelling and to segment transit travelers using smart card data. The travel iterations of individual traveler who use smart card data can be find out Decision tree algorithm is used to mine daily and persistent travelers. Finally segmentation of passengers is done into Stable travelers, Daily Traveler, Persistent traveler, Occasional traveler. A priori market segmentation is used for this, an interesting pattern of users is obtained from each category of passengers.

REFERENCES

[1] Goyat, Sulekha. "The basis of market segmentation: a critical review of literature." *European Journal of Business and Management* 3.9 (2011): 45-54.

[2] Zhang, Fan, et al. "Spatio-Temporal Segmentation of Metro Trips Using Smart Card Data." (2015).

[3] Ma, Xiao-lei, et al. "Transit smart card data mining for passenger origin information extraction." *Journal of Zhejiang University Science C* 13.10 (2012): 750-760.

[4] Harrison, Anna, Vesna Popovic, and Ben Kraal. "A new model for airport passenger segmentation." *Journal of Vacation Marketing* (2015): 1356766715571390.

[5] Chen, Sien, et al. "Understanding Airline Passenger Behavior through PNR, SOW and Webtrends Data Analysis." *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*. IEEE, 2015.

[6] Zhang, Fuzheng, et al. "Reconstructing individual mobility from smart card transactions: a collaborative space alignment approach." *Knowledge and Information Systems* (2014): 1-25.

[7] Kieu, Le Minh, Ashish Bhaskar, and Edward Chung. "Passenger segmentation using smart card data." (2014).

[8] Washington Metropolitan Area Transit Authority (PlantItMetro), Metrorail Data Download, October 2014.