A survey on Travel Pattern Mining

[1]Sona Geeth, [2]Dr. Varghese S. Chooralil
[1]PG Student, [2]Assistant Professor
[1]Computer Science and Engineering,
[1]Rajagiri school of Engineering and Technology, Kochi, India

Abstract: This paper includes the methods to find the travel pattern of the passengers. It is necessary for the transit authorities to identify their customers and provide them with well suited services. Most of the transit authorities uses automatic cash collection system such as smart card instead of ticketing system. This paper gives the information about the passengers who has travelled with a specific origin-destination and with a habitual time. This is a study to find the travel pattern with two clustering techniques. Datas from the transit authorities are used for clustering. Two clustering methods ie; k-means and density based spatial clustering of application with noise(DBSCAN) algorithm are using to cluster the passengers.

IndexTerms - Transit authority, DBSCAN, Clustering.

I. INTRODUCTION

To understand the passengers in a better manner it is important for the transit authorities to fulfill customer needs and preferences. So by clustering the transit agencies would be capable to identify regular origin destination and the habitual time passengers. Transit authorities are the government transportation agencies and they use automatic cash collecting system such as smart cards instead of ticketing system. In spite the large exposure to transit passengers, transit providers have less information about their passengers due to reasons such as the invisibility of passengers, random behavior of the passengers, and the difficulties in validating the disaggregated data of a massive population[1].

The opportunities of the proposed works with the existing are inactive passengers cannot be discovered due to the less information on their mobility requirements and travel behaviors[1]. Existing service improvement projects are limited to the impacts on generic transit customers, avoiding the differences between the types of passengers or segments of passengers with different needs and behaviors.

Clustering is a series of actions by dividing the populations or data points to a fixed groups in a manner that points in a particular group have similar behaviors. In simple, the objective of clustering is to aggregate the groups with similar features and allocate them into clusters. Clustering is the collection of a particular things or data objects based on their features, and aggregates them in accordance with their similarity to their similarities.

In accordance to data mining, this method divides the data enacting a particular join algorithm, more appropriate for the needed information analysis. This clustering methods will allow an object not to be part of a cluster, or strictly belong to it. Density based clustering algorithm plays a significance role in finding non-shapes structure with respect to density. It adopts the idea of density reach ability and density connectivity. Density based algorithms create clusters according to high density of objects of a dataset. It summarizes some distance notion to a density standard level to group members in the clusters. K-means is a simplest unsupervised learning algorithm and it resolves all the clustering problem.

This method results an easy and simple way to aggregate the data points through a fixed number of clusters with a fixed apriori. The main objective is to define K centroids, for all cluster. K means is a particular procedure of vector quantizations particularly from signal processing, which is known for analyzing the cluster in data mining. Cluster analysis is an extremely useful data analysis technique. It has many applications in the science world. All large data set of information can be handled by this kind analysis, resulting great outputs with many varieties of data.

II. CLUSTERING

Clustering is an aggregation of objects which contains in the same class. Otherwise, similar objects will be assigned for one cluster and the objects which are not similar will be assigned in another cluster. Clustering is the series of steps for organizing a group of summary objects into classes of similar objects[15].

1) A cluster of data points can be considered as one group[15].

2) While doing analysis of cluster, we first divide dataset into groups based on similarity of data and then assign the labels to the groups[15].
3) An important advantage of clustering with classification is that, it is flexible to changes and helps single out useful features that identify different groups[15].

Analysis of clustering can be used in market research, recognition of pattern, analyzing data, and image processing. Clustering will also helps to differentiate the documents on the web for discovering the information. As a data mining function, analysis of cluster is a data mining tool which acts as a tool to obtain an accurate awareness into the distribution of data to find the characteristics of each cluster.

2.1 Requirements of Clustering in Data Mining:

1) Scalability - For dealing large datasets we need highly scalable clustering algorithms.
2) Should have the capacity to deal with different kinds of attributes. Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
3) Finding clusters with randomly shape- The clustering algorithm should have the capability to detect clusters of arbitrary shape. They should not be limited to only distance measures that frequently helps to find spherical cluster of with small size.
4) It has the capacity to deal with noisy data that means it will contain noisy, missing or erroneous data. Some algorithms will be sensitive to that kind of data and may cause to make clusters with poor quality.
5) Interpretability The clustering outcomes should be interpretable, comprehensible, and usable.

2.2 Clustering Methods:

1) Partitioning Method
2) Hierarchical Method
3) Density-based Method
4) Grid-Based Method
5) Model-Based Method
6) Constraint-based Method

In partitioning method we will give a set of n objects as data base and partition method will build k partition of data. Each partition will appear as a cluster and k=n. And the constraints are each group has at least one object and each object will belong to completely one group.

Hierarchical method makes a hierarchical decomposition of the given data set. Hierarchical decomposition contains two approaches:(1)Agglomerative approach. It is known as bottom up approach. In this each object will make a different group. And then it will concatenate the objects that are close to one another. And the process will continue until all groups are concatenated into one. (2)Divisive Approach This approach is also known as the top-down approach. In this, clustering can be start with the objects in the same cluster. Through continuous iteration, a cluster will be divided into small cluster. It will be continued until there is an object in each cluster. This method is not adaptable, i.e., once a merging or splitting is done, it can never be undone.

Density-based Method based on the conception of density. The main idea is to continue the clustering process until the density of the neighboring cluster is more than some threshold; ie; for each data point in the cluster , the radius of the cluster should have at least a minimum number of points.

In Grid-based Method the objects together make a grid. The object space is quantized into certain number of cells which makes a grid structure.

In Model-based methods a model is assumed for all cluster to find the best data for the given model. This method will locate the clusters using the density function. It reflects spatial distribution of the data points. This method will give an automatic way to decide the how many clusters are there in accordance with number standard statistics, taking outlier or noise into account. It therefore produce robust clustering methods.

Constraint-based Method: In this method, the clustering is carried out by the combination of user or application oriented constraint. A constraint relates to the expectation of user or the properties of desired clustering results. Constraints gives us with an efficient way of communication with the clustering process. Constraints can be specified by the user or the requirement of application.

III. RESEARCH BACKGROUND

In year 1998, Alexander Hinneburg, Danniel A.kiem proposed an approach to model the overall point density analytically as the sum of influence function of the data points[3]. The advantages of the proposed method are (1) it has a firm mathematical basis, (2)since the proposed method have good clustering features it will work better in datasets with large amounts of noise, (3) it allows a closely mathematical description of clusters with different shape in high- dimensional data sets and (4) it is significantly faster than existing algorithms[3].

In 1996, Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu presented the new clustering algorithm DBSCAN tends on a density-based conception of clusters which is designed to identify clusters of arbitrary shape[14]. DBSCAN only needs only one input parameter and it will support the user in finding an appropriate value for it[14].

In year 2005, Junyi Zhang, Akimasa Fujiwara Makoto Chikaraishi, proposed a method to find the travel patterns and examine their influential factors in the situation of developing countries and a comparison study is organized in order to identify the similarities and differences among travel patterns in different cities [4].

In year 2009, Yu Zheng et al., provide a method by using the GPS trajectories generated by multiple users, we mined interesting locations and classical travel sequences within a given geospatial region[5]. Such information will help us to understand the relation between users and location, and enable travel recommendation as well as mobile tourist guidance. In this work, authors regard an individual's visit to a place as a link from the individual to the location, and weight these links in terms of users travel experiences in various regions [5].

In year 2014 Praveen Rani, Dr., Rajan Vohra Anju Gulia provide a method to discover the travel patterns by the analysis of passport and visa database. This study, gives the information about the choice of destination of Indian passengers, they choose for the fulfillment of their specific purpose[6].

IV. METHODOLOGY

Clustering is the unsupervised classification of patterns (observations, data items, vectors) into groups (clusters) [7]. The cluster groups will form in a manner that data in the same group will be similar and data in other groups will be dissimilar[8]. Clustering has many applications. It includes part family formation for group technology, image segmentation, retrieval of information, summarizing web pages, market segmentation, and scientific and engineering analysis [9]. Clustering will be better, if there is a great similarity within the group and great difference between the groups.

4.1 K-Means

K-means [11] is one of the simplest unsupervised non-hierarchical learning methods among all partitioning based clustering methods. It classifies [12] a given set of n clusters, where k is the number of choices of clusters and it is required in advance. The algorithm will continue repeatedly to allocate each data point to one of K groups based on the features that are given. Based on the similarity in features data points will be clustered.

The output of the K-means clustering algorithm will be:

1) The centroids of the K clusters, which can be used to label new data
2) Labels for the training data (each data point is allocated to a single cluster).

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The Choosing K section means, it can describe how the number can be determined.

A cluster centroid is the aggregation of feature values which define the resulting groups. By evaluating the centroid feature weights we can effectively predict what kind of group each cluster represents. Basically k-means works on distance calculation. For this Euclidean distance is used. Euclidean distance is used to calculate the distance between two data points.

K-means is a repeated process of clustering. It will be continued until it gets the best result. Steps of k-means algorithm is as follows:

1) Give the input.
   It should contains the data set, clustering variables and maximum number of clusters.
2) Assign centroid of the cluster
3) Calculate the Euclidean distance. Euclidean distance is the measures used in k-means algorithm. Euclidean distance between observation and the initial cluster centroids 1 and 2 is calculated. Based on Euclidean distance each point is allocated to each of the cluster based on minimum distance.
4) Then go to the next observation in the dataset and calculate Euclidean distance.

5) Then find the Euclidean distance for the next input, and assign the observation based on minimum Euclidean distance and update the cluster centroids. Continue the steps until all observations are assigned.

4.1.1 Advantages of K-means:

1) It is easy to implement.
2) If variables are large, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.
3) K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular.
4) An instance can change cluster (move to another cluster) when the centroids are recomputed.

4.2 DBSCAN

Density-based algorithms will identify the clusters with high density and clusters with low density. For a given set of points in some space, it will aggregate the points that are tightly packed. It is one of the most common clustering algorithm. We adopt a density-based clustering algorithm because of the following reasons.

1) Density-based algorithms are used to identify clusters of high density and noise of low density. In travel pattern analysis, noise is an anomaly travel pattern that does not follow any regular travel pattern or, in other words, trips that are made randomly. The main goal is to find the clusters (regular pattern) and differentiate it with the anomaly pattern[1].
2) Density-based algorithms will find the clusters with any shape and size. Due to the property of human behavior pattern, a travel pattern can form can also form with any shape and size.
3) Density-based algorithms does not need the initial cores or the number of clusters in advance. This property is also very important for analysis of travel pattern analysis because the number of patterns from an individual passenger is unknown[1].
4) Discretization is an important thing in analysis of travel pattern. Literature survey showed that we need a systematic solution for analyzing the spatial and temporal travel pattern without limiting to stop-to-stop repeated trips and time-window discretization.

Density- based scanning algorithms comprehensively gives a flexible range of high density for each passengers spatial and temporal travel pattern[1]. DBSCAN is 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 and because of the four properties of DBSCAN clustering algorithm which has mentioned previously. DBSCAN algorithm defines the clusters as dense regions. And they are partitioned by regions of a lower point density[1]. The algorithm has two parameters: the maximum density reach distance epsilon(e) and the minimum number of points MinPts. A point can be calculated a core point ic if it has at least MinPts (density) within a radius e, as expressed in:[2]. Mode of Ne(ic) must be greater than or equal to Minpts.

Euclidean distance is used as distance metric. Border point ib is the point which has less points than Minpts within e distance[2]. Noise point is the point which has neither core nor a border point[2].

Steps for DBSCAN algorithm is as follows:

1) select an arbitrary point p.
2) Collect all points which are density reachable from p wrt and MinPts.
3) A cluster will form only if a core point p is formed.

4.2.1 Advantages of DBSCAN:

1) No need to specify the number of clusters in advance.
2) It will works on arbitrarily shaped clusters.
3) Has a concept of noise and is strong to outliers.
4) Needs at least two parameters and it should not adaptable to the ordering of the points in the database.
5) Made for accelerate region queries.
6) minpts and epsilon can be assigned by a domain expert.

V. ANALYSIS

K-means algorithm is easy to perform and the other advantage is its fastness to identify the clusters. But it will works only on well shaped clusters. It is sensitive to outliers and noise. And the number of clusters should be desired in advance. DBSCAN algorithm can make clusters with random shape. It is very strong to noise. And it do not need an apriori k. Density based clustering algorithm plays an efficient role in examining non linear shapes structure based on the density.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. Density reachability and density connectivity are the two main ideas that is mainly used in this algorithm. On the other hand, K-means is one of the simplest unsupervised learning algorithm. And it finds as the most popular clustering algorithm.

The procedure leads a simple and easy way to classify for an available data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. The centroids should be arranged in a crafty way, since different places will cause different results.

VI. CONCLUSION

It is important for the transit authorities to understand their passengers for providing better services to them. So for that it is important to observe the travel pattern. Two algorithms has been used to mine the travel pattern. Using the two clustering algorithms like k-means and DBSCAN, we can cluster the passengers with regular origin destination and habitual time. From the analysis we can conclude that DBSCAN will be more effective in travel pattern mining than k-means algorithm since it can make clusters of different shape and it is very strong to noise.

REFERENCES

[1] Le Minh Kieu, Ashish Bhaskar, Edward Chung, Passenger Segmentation Using Smart Card Data, IEEE Transactions on Intelligent Transportation Systems,Vol.16,No.3,june 2015.
[2] .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. 226231.
[3] Ickjai Lee, Guochen Cai, Kyungmi Lee, Mining Points-of-Interest Association Rules from Geo-tagged Photos.
[4] Junyi Zhang, Akimasa Fujiwara, Makoto Chikaraish, Comparative Analysis Of Travel Patterns In The Developing Cities Based On A Hybrid Model, Journal Of The Eastern Asia Society For Transportation Studies, Vol. 6, Pp. 4333 - 4348, 2005.
[5] Yu Zheng, Lizhu Zhang, Xing Xie, WeiYing Ma, Mining Interesting Locations and Travel Sequences from GPS Trajectories,Microsoft Research Asia, 4F, Sigma building, No.49 Zhichun road, Haidian District, Beijing 100190, China.
[6] Praveen Rani, Dr., Rajan Vohra Anju Gulia, Discovering Travel Pattern In Passport Data Analysis, IJIRS, in communicating.
[7] Usama M. Fayyad. Data mining and knowledge discovery: Making sense out of data, IEEE Expert: Intelligent Systems and Their Applications, 11(5):2025, 1996.
[8] Dr. Sankar Rajagopal, Customer Data Clustering Using Data Mining Technique International Journal of Database Management Systems (IJDMS) Vol.3, No.4, November 2011.
[9] Pham, D.T. and Afify, A.A. (2006) Clustering techniques and their applications in engineering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science.
[10] Er. Arpit Gupta, Er. Ankit Gupta, Er. Amit Mishra, Research Paper on Cluster Techniques of Data Variations, International Journal of Advance Technology Engineering Research (IJATER).
[11] Hartigan, J., A. and Wong, M., A. 1979, A K-Means Clustering Algorithm, Applied Statistics, Vol. 28, No. 1, pp. 100-108.
[12] Selim, S., Z. and Ismail, M., A. 1984, K- Means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality, IEEE Trans. Pattern Anal. Mach. Intel., Vol. 6, No. 1, pp. 8187.
[13] Hinneburg and D. A. Keim, An efficient approach to clustering in large multimedia databases with noise, in Proc. KDD, 1998, pp. 5865.
[14]M. Ester, H.-P.Kriegel, J. Sander, and X. Xu, A density-based algorithm fordiscoveringclustersinlargespatialdatabaseswithnoise,inProc.2nd Int. Conf. Knowl. Discov. Data Mining, 1996, pp. 226231.
[15] https://www.tutorialspoint.com/datamining/dmclusteranalysis.htmL