Hierarchical Data Clustering Model for Analyzing Passengers Trip in Highways

S. O. Hasanpour Jesri, A. Ahmadi, B. Karimi* & M. Akbarpour Shirazi

Seyed Omid Hasanpour Jesri, M.Sc. of Science, Department of Industrial Engineering & Management Systems, Amirkabir University of Technology, Tehran, Iran, Abbas Ahmadi, Assistant Professor, Department of Industrial Engineering & Management Systems, Amirkabir University of Technology, Tehran, Behrooz Karimi, Associate Professor, Department of Industrial Engineering & Management Systems, Amirkabir University of Technology, Tehran, Mohsen Akbarpour Shirazi, Assistant Professor, Department of Industrial Engineering & Management Systems, Amirkabir University of Technology, Tehran, Iran.

KEYWORDS
Urban planning-
data mining-
traffic analysis-trip-
public transportation

ABSTRACT
One of the most important issues in urban planning is developing sustainable public transportation. The basic condition for this purpose is analyzing current condition especially based on data. Data mining is a set of new techniques that are beyond statistical data analyzing. Clustering techniques is a subset of it that one of it’s techniques used for analyzing passengers’ trip. The result of this research shows relations and similarities in different segments that its usage is from strategic to tactical and operational areas. The approach in transportation is completely novel in the part of trip patterns and a novel process is proposed that can be implemented in highway analysis. Also this method can be applied in traffic and vehicle treats that need automatic number plate recognition (ANPR) for data gathering. A real case study has been studied here by developed process.

© 2012 IUST Publication, IJIEPR, Vol. 23, No. 4, All Rights Reserved.

1. Introduction
This paper intended to analyze traffic in highways. This approach in transportation in the part of trip patterns is completely novel and a novel process is proposed that can be implemented in highway analysis. The results of this paper can be generally used in the field of urban planning especially in public transportation scheduling and allocation of fleets such as bus and taxi.

No similar research has been done for analyzing traffic and trip patterns by data mining approaches; however data mining methods have been used in location and transportation problems for other purposes that will be presented in the literature review. Data mining techniques have been used broadly to tackle different engineering problems such as document retrieval, image segmentation, speech recognition and bioinformatics and many other fields even in medical science [18, 19, 20, 21, 22, 23, 24].

At the first, the traditional models in analyzing trip patterns and trip distribution are introduced and then the differences between them and the proposed model are described. After that the basic definition of clustering is presented and then, in the literature review section, the application of data mining in transportation issues will be addressed.

The traditional models used for forecasting and trip patterns analysis consist of: 1) trip generation models including Growth Factor Models, Multiple Regression Models, Cross Classification Model. 2) Trip distribution models such as Average Growth Factor Model, Detroit Growth Factor Model, Frater Model, Gravity Model and Linear Programming Method [25]. The main differences between the suggested model and traditional models are as follows:

* Corresponding author: Behrooz Karimi
Email: b.karimi@aut.ac.ir
Paper first received July. 05, 2012, and in revised form Oct. 9, 2012.
1) In the traditional models the results are based on approximations, but here, raw and pure data is used in the proposed model. As a result, the proposed model is completely examined with the real data.

2) In all of the traditional models, primary information needed is usually presented in large scale and is not easily accessible but in the proposed model, raw data is gathered in detail.

3) The proposed approach is capable of handling changes and different conditions while traditional models cannot be updated in short periods. Also, the proposed model is extendable to all kind of highways.

Therefore, a new approach in trip patterns and passenger’s behavior is used that is based on data mining techniques especially clustering. Data clustering is to group objects so that one group’s objects be similar or connected and objects in different groups be different or unrelated [1, 2]. Clustering is generally categorized into 2 main approaches: partition clustering and hierarchical clustering. In the first one, the number of clusters is defined as a parameter but in hierarchical clustering, the number of clusters is determined by the similarity that user has selected in advance as a parameter. Hierarchical clustering itself is divided to 2 categories; Divisive and Agglomerative. Divisive clustering puts all objects in one cluster at first. Then by considering acceptable distances, the first cluster is divided to two clusters that satisfy the distance parameter. This procedure would continue to obtain clusters as many as objects. In other words, at the end of this approach, the number of clusters will be equal to the number of objects. Agglomerative clustering is reverse form of the divisive methods [3].

2. Literature Review

Clustering techniques are used in allocation, Routing, Location - allocation problems and routing - location problems. In location and routing problems, Jain and Dubes defined the cluster as joint regions from a multidimensional space involved by more density places which has divided by a region with rather little density, from other similar regions [4]. Moreover MFLP (Multiple Facility Location Problem) is a specific clustering problem, where the customers are allocated to a resource. Since the expenses are assumed by blow overed distance in MFLP, every customer is allocated to the nearest resource in optimize conditions. “Bar and Ball” model discovers some places in graph or network where the traveler in their way to destination are temporarily stopped. Hub is defined by Campbell: “Facilities as replacement and exchange point, which serve origins and destinations in transportation systems” [7,8]. Hub location problem is a kind of location-allocation problem so it is considered as a clustering problem [9].

One of the research areas in data mining which is almost common in transportation is clustering arcs and nodes in transportation network problem. In this kind of problem, transportation network is considered as a graph. The properties of nodes involve: axis of coordinates(x, y), arrived and departed link to each node, the situation of traffic jam (like pilot and barred crossing symptoms). The properties of arcs involve: the places of origin and destination ways, the number of lines, the peak of the speed, the length of road, the speed of current possible trip with regard to traffic jam. Then nodes and lines are clustered with their traits, to be able to derivate the suitable queries in necessary time [10, 11].

Considerable efforts have been performed to group and cluster subway stations. The major efforts actually concentrate on clustering the stations of a subway, so that the main attribute of this clustering is frequency of travelers in each station and this data is used for strategic transportation and land use planning. In fact the stations are studied considered insular and the relations have not been purposed [12, 13].

Most of the efforts which have been performed on transportation networks are based on statistical methods due to analysis the structure of networks. One of these researches is a statistical analysis on subway in Seoul (a city in South Korea) and the flow of travelers where it has 380 stations [14]. There are a lot of methods for analyzing the transportation issues; otherwise there are two main methods which are efficient: statistical methods and neural network methods. Vlahogianni has presented a comparison between these 2 methods and the advantages and disadvantages of these methods in analyzing transportation data [15]. In a different analysis, a new method of planning for designing a road system is presented. In this model plenty of variables are considered based on an AHP model [16]. After literature review, we are going to illustrate the problem. Here, the treats of traffic jam and travelers in different hours in a highway are surveyed. Therefore, the highway should be divided to smaller segments. Which has similar treats in aspect of traffic travelers, it means that the variations of purpose variable would not be sensible in each part. Sampling methods are often used for data gathering. Since in designing the questionnaire, the data mining aspect is not usually considered, therefore the data mining approach is up to downward or undirected, it means that there is not any special knowledge for derivation and the purpose was mute iterative and interactive data and the knowledge is derived in regard to nature of gathered data. However, if we notice to the data mining objectives in data gathering step, it would be easier to access the knowledge.

This objective has some advantages such as, no cost should be paid for data gathering; In fact we can use the data which are collected as a matter, in spite of iterative and interactive process without any expenses.
3. Modeling

Data mining is based on a process developed by Dubes and Jain has a basic structure. The basis of current paper is similar to this structure to some extent. A special-purpose process for mining the travel data is presented that is the main contribution of this paper. In this model number of passengers in each period is considered as a case that looking for finding same patterns in passenger trips in different sections and periods, so the model presented here, first clusters time periods as variables, then in each cluster, it tries to find similar segments to identify patterns and then it tracks the behavior of them in other time periods. The steps of model are presented in the following:

1) Data collection
2) Data preparation and extracting target data
3) Determining clustering strategy
4) Clustering variables (time periods)

Here variables are time intervals; it means that the data of any time period is considered as a variable. In this step the objective is to find similar periods of time in each day that passengers’ patterns are almost the same. For clarifying this issue an example will be provided. In traditional grouping, the Horizon between 6:30 AM to 8:00 AM is usually considered as a peak traffic hour and the same scheduling and allocation program is considered there for public transportation. Here this assumption is taken into consideration.

5) Clustering segments in any time period clusters

After identification of time clusters, segments in each cluster which are considered as objects are examined to understand which parts of highway have similar treat in terms of passenger volume in any time period. In the previous studies, there exists no similar analysis, as per our knowledge, and the connection and linkage between different segments have not been addressed. However, the proposed approach shows there are too many powerful and meaningful relationships between different parts of highway and passengers’ trip patterns.

6) Finding patterns

After clustering time periods and segments, patterns and their properties should be found. Indeed in this step, this assumption is examined that whether these patterns are generated randomly or are frequent and repetitive.

7) Pattern analysis

Now patterns are analyzed to understand the main properties and attributes such as starting and finishing times of each pattern, change in patterns and movements.

8) Knowledge extraction

After recognizing patterns, knowledge related to the objective of project can be easily extracted from meaningful patterns. This knowledge is used in variety of areas from operational to tactical and also strategic planning.

4. Case Study: Resalat Highway

Resalat highway is one of the main and longest highways in Tehran. This highway connects the west and east area of Tehran together. Here, the clustering methods are used for discovering knowledge about patterns of passenger’s urban trip in this highway. The purpose of this study is to collect knowledge for designing Bus Rapid Transit (BRT) in this highway. In data gathering step, the highway has been divided to ten sections in both directions: west to east and the opposite way. This data includes number of Taxies and cars and number of passengers in each vehicle that is counted in any segment each 30 minutes from 6 AM until 8:00 PM. For the sake of simplification in presentation of the Resalat highway, this highway and related segments are shown in Fig. 1 in which odd segments show East to West direction and even segments show West to East direction and segments 1 and 20 are the starting segments in East to West and West to East direction, respectively.

Now the proposed model is applied in this case according to the steps aforementioned:

1) Data gathering

The data was gathered in 2nd and 3rd of May2011 by observation and was recorded in the forms for analyzing. The primary objective of the data gathering step was the simulation of the new scenario (considering BRT alternatives) but simulation could not recognize patterns and their change in temporal data. Thus, clustering approach was taken into account and it requires no cost to be paid for data gathering in data mining process.

2) Data preparation and preprocess

As the object of data mining is to discover trip patterns of passengers so the data related to volume of vehicles is useless, and the summation of passengers in all vehicles should be extracted.

3) Determining clustering strategy

Here the model of clustering should be defined. As it was mentioned, there are two main groups of clustering techniques: partitional and hierarchical ones. In this case, hierarchical clustering is used because it is more flexible than partitional one and it has this ability to insert human intelligence in determining the parameters of model. Theagglomerative hierarchical...
clustering was used the Squared Euclidean distance was used to calculate the distances and finally the kind of merging used is “between group linkages” method.

4) Clustering variables (time periods)
As mentioned in the model description, first we should understand which hours are similar together from passenger’s trip pattern. Here the section of time horizon is half hour; it means that the data for each 30 minutes from 6 AM until 8 PM was recorded. Thus, there were 28 primary periods. For clustering, the SPSS Statistics software version 17 was used. According to the acceptable distance that has been set as a parameter, 28 primary periods are clustered in 13 new periods that were much similar. These clusters are: [6 AM-6:30 Am]- [6:30 AM-7 AM]- [7 AM-7:30 AM]- [7:30 AM- 8 AM]- [8 AM-8:30 AM]- [8:30 AM-10:30 AM]- [10:30 AM-12 AM]- [12 AM- 1:30 PM]- [1:30 PM- 3 PM]- [3 PM- 4 PM]- [4 PM- 5 PM]- [5 PM- 7 PM]- [7 PM- 8 PM].

The dendrogram of time clustering is shown in figure 2. The result shows trip and traffic patterns are not the same in rush hour and it has large fluctuations due to change in volume of passenger trips.

![Dendrogram of time variable clustering](image)

**Fig. 2.** Dendrogram of time variable clustering

5) Clustering segments in any time period clusters
In this step the similar segments in each time period cluster are extracted. By the same clustering approach used in previous step, in each time clusters, the segments are clustered according to passenger’s trip behavior. For example the dendrogram of clustering in period 4 PM to 5 PM is shown in Fig.3.

![Dendrogram of clustering segments in 4 PM to 5 PM](image)

**Fig. 3.** Dendrogram of clustering segments in 4 PM to 5 PM

The result of clustering is shown schematically in Table I. Each segment is shown with a color and the segments have the same color belong to the same cluster and the segments which have no color were put in individual cluster in this step. For example segments 1, 3, 7, 2 and 4 are in the same cluster.

**Tab. 1.** Segments clustering result at 3PM to 4PM

<table>
<thead>
<tr>
<th>Segment number</th>
<th>direction</th>
<th>result at 3PM to 4PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>East To West</td>
<td>58</td>
</tr>
<tr>
<td>3</td>
<td>East To West</td>
<td>219</td>
</tr>
<tr>
<td>5</td>
<td>East To West</td>
<td>578</td>
</tr>
<tr>
<td>7</td>
<td>East To West</td>
<td>202</td>
</tr>
<tr>
<td>9</td>
<td>East To West</td>
<td>400</td>
</tr>
<tr>
<td>11</td>
<td>East To West</td>
<td>2109</td>
</tr>
<tr>
<td>13</td>
<td>East To West</td>
<td>987</td>
</tr>
<tr>
<td>15</td>
<td>East To West</td>
<td>1716</td>
</tr>
<tr>
<td>17</td>
<td>East To West</td>
<td>976</td>
</tr>
<tr>
<td>19</td>
<td>East To West</td>
<td>824</td>
</tr>
<tr>
<td>2</td>
<td>West to East</td>
<td>72</td>
</tr>
<tr>
<td>4</td>
<td>West to East</td>
<td>191</td>
</tr>
<tr>
<td>6</td>
<td>West to East</td>
<td>277</td>
</tr>
<tr>
<td>8</td>
<td>West to East</td>
<td>352</td>
</tr>
<tr>
<td>10</td>
<td>West to East</td>
<td>678</td>
</tr>
<tr>
<td>12</td>
<td>West to East</td>
<td>1314</td>
</tr>
<tr>
<td>14</td>
<td>West to East</td>
<td>992</td>
</tr>
<tr>
<td>16</td>
<td>West to East</td>
<td>1039</td>
</tr>
<tr>
<td>18</td>
<td>West to East</td>
<td>754</td>
</tr>
<tr>
<td>20</td>
<td>West to East</td>
<td>801</td>
</tr>
</tbody>
</table>

International Journal of Industrial Engineering & Production Research, December 2012, Vol. 23, No. 4
6) Finding patterns
The patterns created randomly or have high frequency in other time periods and are meaningful have to be checked. If we do clustering for all segments in all periods, we can then judge about this problem. Table II shows the patterns in this horizon from 6AM to 12AM, it is obvious that most patterns have high frequency and are meaningful and are not generated randomly.

7) Pattern analysis
Now, we should specify the properties of main patterns recognized. The patterns are called main if they have at least one of the following conditions: 1) the patterns have long periods and high frequency; 2) the patterns include large number of segments. In Table III main patterns and some attributes of them are mentioned.

Tab. 2. Patterns in horizon from 6AM to 12AM

<table>
<thead>
<tr>
<th>Segment number</th>
<th>Status of passenger traffic (people)</th>
<th>Hours of pattern formation</th>
</tr>
</thead>
<tbody>
<tr>
<td>East to West</td>
<td>Less than 200</td>
<td>6AM to 10:30AM</td>
</tr>
<tr>
<td>9, 5</td>
<td>About 250</td>
<td>6AM to 3PM</td>
</tr>
<tr>
<td>9, 13, 6</td>
<td>About 400</td>
<td>3PM to 5PM</td>
</tr>
<tr>
<td>5, 13, 19</td>
<td>About 575</td>
<td>5PM to 7PM</td>
</tr>
<tr>
<td>5, 19, 8</td>
<td>About 650</td>
<td>4PM to 5PM</td>
</tr>
<tr>
<td>1, 2, 3, 9</td>
<td>About 200</td>
<td>8:30AM to 5PM</td>
</tr>
<tr>
<td>2, 3, 4</td>
<td>Less than 200</td>
<td>5PM to 7PM</td>
</tr>
<tr>
<td>11, 19</td>
<td>Upward trend from 800 to 1400</td>
<td>6AM to 7:30AM</td>
</tr>
<tr>
<td>17, 19</td>
<td>Downward trend from about 1000 to about 800</td>
<td>8:30 to 10:30AM</td>
</tr>
<tr>
<td>6, 8</td>
<td>Downward trend from about 1000 to about 800</td>
<td>6:30 to 10AM</td>
</tr>
<tr>
<td>12, 14</td>
<td>First ascending from 800 to 1400 then descending from about 1000 to below 1000</td>
<td>6:30AM to 7:30AM and 8AM to 8:30AM</td>
</tr>
<tr>
<td>10, 14</td>
<td>About 1100</td>
<td>6:30AM to 10:30AM</td>
</tr>
<tr>
<td>8, 10, 19, 20</td>
<td>About 650</td>
<td>10:30AM to 12AM</td>
</tr>
<tr>
<td>16, 17</td>
<td>About 900</td>
<td>10:30 to 4PM</td>
</tr>
<tr>
<td>18, 20</td>
<td>Rising from 600 to more than 1300</td>
<td>12AM to 4PM</td>
</tr>
<tr>
<td>19, 20</td>
<td>About 700</td>
<td>10:30AM to 4PM</td>
</tr>
<tr>
<td>14, 16, 17</td>
<td>About 1000</td>
<td>12AM to 3PM</td>
</tr>
<tr>
<td>12, 19</td>
<td>Descending from 1200 to 900</td>
<td>5PM to 7PM</td>
</tr>
<tr>
<td>5, 9, 6</td>
<td>About 350</td>
<td>12AM to 3PM</td>
</tr>
</tbody>
</table>

8) Knowledge extraction
In the final step according to the patterns recognized, the knowledge should be discovered for objective purposes. Some applications can be used here from patterns recognized are: designing the stations includes Station dimensions and corridor number in accordance with the prevailing patterns, determining the optimum number of personnel with regard to passenger traffic patterns and also exchanging of forces between the stations due to the changing patterns in different hours,
scheduling and balancing and headway determination for stations and Regulating passenger traffic signals proportional to the density of the patterns.

5. Conclusions

This paper shows that data mining techniques such as clustering can be useful in transportation analysis and it presents a novel approach in transportation problems. The traditional models used in distribution modeling of passenger’s trip can be revised and updated by this approach. The model presented in this paper is applicable to all kinds of highways. The result is very valuable by considering the fact that no cost is paid for data gathering and its application spreads from strategic to tactical and operational areas, as the outputs of this paper can be used in designing bus line (strategic area), fleet allocation (tactical area) and scheduling and headway determination (operational area).

For further research, we suggest to apply data mining techniques real time by using geographic analysis and looking for the root of patterns such as land use effects and population distribution.

References


Hierarchical Data Clustering