International Journal of Industrial Engineering & Production Research

December 2012, Volume 23, Number 4
pp. 253-259

http://IJIEPR.iust.ac.ir/

ISSN: 2008-4889

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.

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 sfor 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 purposed 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.

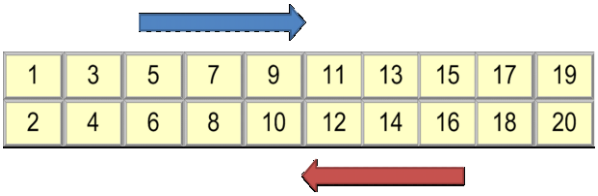

Fig. 1. Segmentation of Reaslat highway

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

clusteringwas 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.

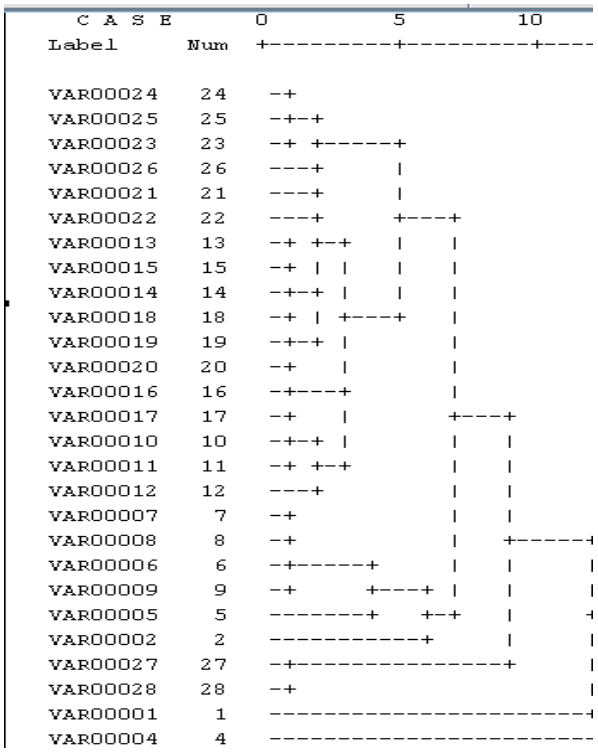

Fig. 2.Dendrogram of time variable clustering

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.

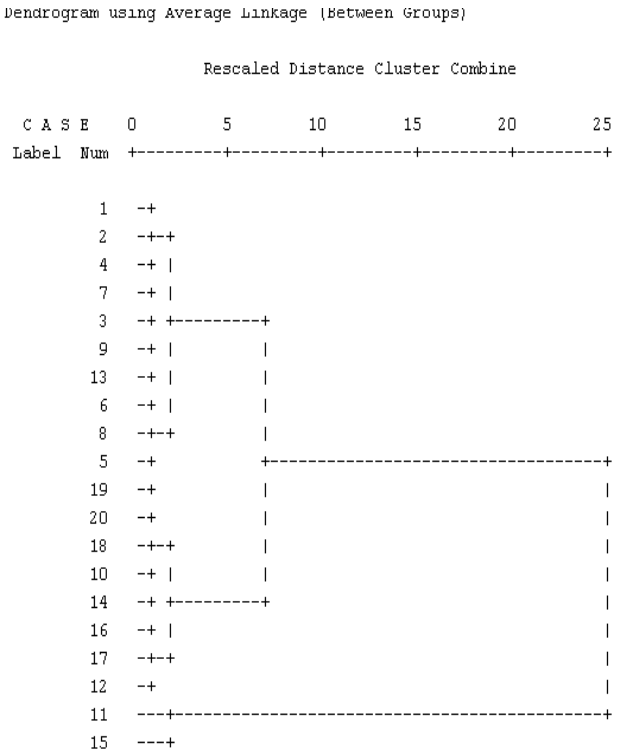

Fig. 3. Dendrogram of clustering segments in 4 PM to 5 PM

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

Segments Clustering	result at 3PM to 4PM		
segment number	direction	3PM-3:30PM	3:30 PM-4PM
1	East To West	58	81
3	East To West	219	117
5	East To West	578	525
7	East To West	202	195
9	East To West	400	381
11	East To West	2109	2061
13	East To West	397	421
15	East To West	1716	1814
17	East To West	976	980
19	East To West	824	746
2	West to East	72	85
4	West to East	191	158
6	West to East	277	440
8	West to East	352	616
10	West to East	678	961
12	West to East	1314	1196
14	West to East	992	1116
16	West to East	1039	1115
18	West to East	754	758
20	West to East	801	756

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

Segment number		6/00-6/30	6/30-7/00	7/00-7/30	7/30-8/00	8/00-8/30	8/30-9/00	9/00-9/30	9/30-10/0	0/00-10/3	0/30-11/0	1/00-11/3	1/30-12/0
1	East to West	40	129	176	127	112	152	126	189	154	148	203	140
3	East to West	52	140	179	141	89	84	115	127	78	93	65	76
5	East to West	94	223	262	278	148	172	172	181	278	262	320	295
7	East to West	75	106	89	109	80	90	92	80	129	175	128	151
9	East to West	82	192	184	205	163	164	175	191	239	229	257	325
11	East to West	884	1222	1275	1586	1689	1588	1411	1585	1585	1682	1750	1769
13	East to West	427	670	730	732	711	744	825	708	695	560	527	658
15	East to West	1228	1928	3721	1870	2382	1653	1703	1544	1389	1523	1483	993
17	East to West	481	838	870	791	1046	1100	995	883	900	823	726	809
19	East to West	840	1214	1421	2252	976	937	725	768	729	702	719	556
2	West to East	162	272	428	299	229	228	154	150	144	147	118	98
4	West to East	240	511	674	454	421	357	319	322	395	242	274	227
6	West to East	353	1015	722	876	742	840	626	698	477	388	455	432
8	West to East	641	968	549	1001	696	892	787	712	855	680	770	596
10	West to East	427	697	799	1070	1172	1125	999	1124	1071	743	636	711
12	West to East	758	790	1087	1253	1443	1330	1348	1490	1448	1371	1122	1230
14	West to East	571	741	1033	981	1351	1130	1195	1077	1115	1112	1015	964
16	West to East	425	866	795	569	475	663	578	668	907	772	835	785
18	West to East	211	291	361	486	404	432	433	487	563	500	520	732
20	West to East	631	918	848	839	694	705	693	822	599	740	611	650

Tab. 3. Attributes of main patterns

Patterns Properties	Main Attributes	
Similar Segments	*Status of passenger traffic (people)*	*Hours of pattern formation*
1, 3, 5, 7, 9	Less than 200	6AM to 10:30AM
5, 9	About 250	6AM to 3PM
9, 13, 6	About 400	3PM to 5PM
5, 13, 19	About 550	5PM to 7PM
5, 19, 8	About 650	4PM to 5PM
1, 2, 3, 9	About 200	8:30AM to 5PM
2, 3, 4	Less than 200	5PM to 7PM
11, 19	Upward trend from 800 to 1400	6AM to 7:30AM
17, 19	Downward trend from about 1000 to about 800	8:30 to 10:30AM
6, 8	Downward trend from about 1000 to 800	6:30 to 10AM
12, 14	First ascending from 800 to 1400 then descending from 1400 to below 1000	6:30AM to 7:30AM and 8AM to 8:30AM and 12AM to 1:30PM
10, 14	About 1100	8:30AM to 10:30AM
8,10,19,20	About 650	10:30AM to 12AM
16, 17	About 900	10:30 to 4PM
18, 20	Rising from 600 to more than 1300	12AM to 8PM (except 5PM to 6:30PM)
19, 20	About 700	10:30AM to 4PM
14, 16, 17	About 1000	12AM to 3PM
12, 19	descending from 1200 to 900	5PM to 7PM
5, 9, 6	About 350	12AM to 3PM

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 stationsand 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

[1] Hansen, P., Jaumard, B., *"Cluster Analysis and Mathematical Programming*". Mathematical programming, 1979.

[2] Anderberg, M.R., "*Cluster Analysis for Applications*". Academic Press, New York, 1973.

[3] Jiawei Han, Michelle Kamber, *"Data Mining: Concepts and Techniques*". Morgan Kaufmann, 2001, pp. 443-495.

[4] Jain, A.K., Dubes, R.C., *Algorithms for Clustering Data". Prentice Hall, Englewood Clifs*, New Jersey,1988.

[5] Cooper, L., *"Location-Allocation Problem*". Operation Research, 1963.

[6] Cooper, L., "*Heuristic Methods for Location-Allocation Problems*". SIAM Review, 1964.

[7] Campbell, J.C., Radke, J., Gless, J.T., Wirtshafter, R.M., "*An Application of Linear Programming and Geographic Information Systems: Cropland Allocation in Antigua*". Environment and planning A, 1992.

[8] O'Kelly, M.E., *"A Clustering Approach to the Planar Hub Location Problem*". Annals of operation Research 40, 1992b.

[9] Bernd Wagner. *"An Exact Solution Procedure for a Cluster Hub Location Problem*", European Journal of Operation Research, 2007.

[10] Yun Wu Huang, Ning Jing, Elke, A.,*"Optimizing Path Query Performance: Graph Clustering Strategies*", Rundensteiner.IBM T.J.Watson Research Center.2000.

[11] Agrawal, R., Kiernan, J., *"An Access Structure for Generalized Transitive Closure Queries*". IEEE Ninth International Conference on Data Engineering.1993.

[12] Stefan Zemp, Michael Stauffacher, Daniel J., Lang, Roland, W., Scholz, *"Classifying Railway Stations for Strategic Transport and Land use Planning: Context Matters",* Journal of Transport Geography.2010.

[13] Verhetsel, A., Vanelslander."*What Location Policy Can Bring to Sustainablecommuting: an Empirical Study in Brussels and Flanders",* Journal of Transport Geography.2010.

[14] Keumsook Lee,Woo-Sung Jung , Jong Soo Park , M.Y. Choi."*Statistical Analysis of the Metropolitan Seoul Subway System: Network Structure and Passenger Flows",* Center for Transportation Studies, Boston University.2008.

[15] Karlaftis, M.G., Vlahogianni, E.I., "*Statistical Methods Versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights*" Department of Transportation Planning and Engineering, School of Civil Engineering, National Technical University of Athens, Greece. 2011.

[16] Abolghasem Sadeghi-Niaraki , Masood Varshosaz, Kyehyun Kim, Jason J. Jung.*"Real World Representation of a Road Network for Route Planning in GIS. Dept. of Geoinformatic Eng*", Inha Univ., Incheon, South Korea.2011.

[17] Anil, K., Jain and Richard C., Dubes*." Algorithms for Clustering Data"*. Prentice-Hall, 1988.

[18] Jain AK, Murty MN, Flynn PJ. "Data clustering: a review:, ACM Comput Surv 31(3):264–323, 1999.

[19] Ahmadi, A., Karray, F., Kamel, M.S., *"Multiple Cooperating Swarms for Data Clustering. In: Proceeding of the IEEE Swarm Intelligence Symposium*", Honolulu, Hawaii, 1–5 April 2007.

[20] Ahmadi, A., Karray, F., Kamel, M.S., "*Particle Swarm-Based Approaches for Clustering Phoneme Data*", In:UW and IEEE Kitchener-Waterloo section joint workshop on multimedia mining and knowledge discovery, Waterloo, Canada, 17–18 Oct, 2007.

[21] Cui, X., Potok, T.E., Palathingal, P., *"Document Clustering using Particle Swarm Optimization*", In:Proceeding of the IEEE swarm intelligence symposium, Pasadena, CA, 8–10 June 2005.

[22] Omran, M., Engelbrecht, A.P., Salman, "*A Particle Swarm Optimization Method for Image Clustering*". Int J Pattern Recognit Artif Intell 19(3): 2005, 297–321.

[23] Omran, M., Salman, A., Engelbrecht, A.P., *"Dynamic Clustering using Particle Swarm Optimization with Application in Image Segmentation*", Pattern Anal Appl 6: 2006, pp. 332–344.

[24] Xiao, X., Dow, E.R., Eberhart, R., Miled, Z.B,, Oppelt, R.J., "*Gene Clustering using Self-Organizing Maps and*

Particle Swarm Optimization", In: Proceeding of international parallel processing symposium, Nice, France, 22–26 Apr 2003.

[25] Khisti, C., Jotin, *"Transportation Engineering: an Introduction"*, Prentice Hall, 2nd editon,1998.