Clustering of Public Transport Operation using K-Means

Rabiah Abdul Kadir
Institute of Visual Informatics (IVI), UKM 43600 Bangi , Selangor , Malaysia
rabiahivi@ukm.edu.my

Riza Sulaiman
Institute of Visual Informatics (IVI), UKM 43600 Bangi , Selangor , Malaysia
riza@ukm.edu.my

Yasuki Shima
Institute of Visual Informatics (IVI), UKM 43600 Bangi , Selangor , Malaysia
y.shima@mail.meio-u.ac.jp

Fathelalem Ali
Faculty of International Studies, Meio University 1220-1 Bimata, Nago, Okinawa 905-8585, Japan
ali@mail.meio-u.ac.jp

***Abstract*—This paper describes a methodology for analysing the operation of transport buses, using a simplified generic approach. In recent years, location systems that utilize GPS data have become widespread, including their use for monitoring the operation of buses. In this paper, we simplify the bus routes, monitoring process, present new insights using K-means, and enhance their effectiveness. The primary focus of this work is data collection, subsequent data analysis and reporting, for use in schedule adjustments and resource allocation. The experimental data for this study is obtained from the operation of public transport buses at the main campus of Universiti Kebangsaan Malaysia (UKM), where several buses operate on a 30-minute interval over three different concurrent routes. The proposed approach to data analysis is based on three attributes of the data being collected, namely time, volume and quality. The effectiveness of the proposed approach described and discussed.**

Keywords-Internet of Things; Big Data Analysis; GPS; Visualisation; K-means

I. INTRODUCTION

Along with the fact that huge budgets are allocated for developing transportation systems, but still in many cities and town roads, efficient performance of transportation systems is questioned. For a variety of reasons, the entangled transportation systems cannot satisfy the numerous requirements of urban mobility. In case of Malaysia, even though the Klang Valley area of the Malaysia Peninsula is developed with the intercity transportation such as ERL, Commuter, LRT, and monorail take in place (Mohamed, 2012). However, the travelling between the Kuala Lumpur International Airport (KLIA) and the centre of Kuala Lumpur by bus is still in hassle with the traffic and inconvenient such as in Cyberjaya (Natsumi, 2013). Some earlier research work stated that the main problem of the bus services usually refers to the bus arrival that always delayed from the actual schedule (Suwardo, 2010). Information regarding the bus schedule, timetable and routes are only located at the bus stop or bus station (Mohammad, 2011; Tsutomu, 2006), and still no much real time information available for passengers and operators.

Many are promoting the use of big data analysis in resolving issues to transportation. The objective of big data analysis is to use the large volume of data to extract new knowledge by searching, for example, for patterns in the data (Gavin, 2015). However, not many companies really need to go to big data and its analysis tools to optimize their interests (Jeanne, 2013). In this paper, we present a framework for simplified ways for collecting and analysing data for bus operation monitoring.

II. RELATED WORK

The effectiveness of the bus operation is a core research for the public transport. There are several research have been done on this problem and it was studied in two different points of view, first was focused on the human factor in giving hospitality to the passengers and secondly, was in the processing of the data collected. Research on human factor is on facilities and hospitality provided to the passenger such as the quality of services and the role of driver in assisting the passengers (Munzilah, 2013). Whereas on data processing, a tool such as Advanced Traveller Information System (ATIS) was used to monitor and maintain most up-to-date information of public transport. Data were collected using Automatic Passenger Counter (APC) on buses operated by a transit agent. Data mining tools, CART (Classification and Regression Trees) for decision trees and SAS (Version 8.02) for Hierarchical Clustering were used to analyse the data and extract information (Jayakrishna, 2006).

In collecting data, automated data collection holds the key to doing statistically valid analyses of running time and schedule adherence (Peter G, 2000). Many of the recent active research efforts uses Global Positioning System (GPS) to gather the information about public transportation and the congested places especially in urban area (Andrei, 2010). GPS is commonly used to track information about location and time in all conditions.

GPS is a common tool to collect data in real time. Leon et al (in Leon, 2013) studied the data that are collected from GPS to derive bus stop locations, route geometries and service schedule, without the need for information from transit agencies for public transport such as buses and trains. They evaluated the classification features with spatial and temporal clustering techniques to update route geometries, transit stop locations or service schedules. Other researchers tried measuring the arrival time of the bus by using GPS to collect the data (S. L. Bangare, 2013).

978-1-5386-4794-3/18/$31.00 ©2018 IEEE

Recent years also are witnessing many research and applications using intelligent systems and big data analysis in transport system monitoring operation optimization (Gang Zeng, 2015). In contrast to big data analysis approaches that focus on real time big data properties that mostly described by the 5V's model (H. V. Jagadish, 2014): Volume, Velocity, Variety, Veracity, Value, we use a 3-property approach to monitor transport system, vehicle operations, and that can easy to use for small and medium sized companies in particular.

In this study, we use sports tracker application to gather the appropriate information such as the movement of the public bus along the route of the bus. The data collected showed the movement of the bus from one location to another location and it is significant to be implemented in monitoring the public bus operation. Data analysis is based on three attributes of data being collected, namely, time, volume and quality.

III. Proposed Framework

Our approach describes three steps; collection of data, analysis and visualisation. The data organized in regard with three attributes; time, volume, and quality. The analysis and visualisation are based on those three attributes. Following the step, a clustering non-supervised algorithm (K-means Clustering (MacQueen, 1967)) is used as a method of analysing those three attributes in the data being collected.

- **Time attribute:** Is the time of occurrence of measures or incidents that resembles the core of data being available or under consideration. The time, typically measurable either automatically using tools and devices or can be recorded manually.
- **Volume attribute:** This refers to the amount of data available; it may resemble the collection of iterations of incidents that share common attributes but may differ in some attributes such as time.
- **Quality attributes:** This is defined in terms of interest of customer or service provider. It comprises the most credible part of data that related to need of consideration of customer or service provider. It can be available in the basic dataset, but also can be availed using some mathematical or analytical process, from the basic data set. The quality attributes are compared to Value of big data as described in some of works related to big data analysis (H. V. Jagadish, 2014).

Quality evaluation: Information (time, latitude, longitude) obtained from the GPS receiver can grasp the distance moved. When data obtained are classified by the unsupervised learning algorithm (K-means), move tendency and stops tendency can be classified. The analysis result obtained from these migration logs confirms what value can be obtained from the bus company. The conventional system is focused on knowing the distance travelled, the location of movement, from the GPS data, however, this research discovers new values from different perspectives by applying the K-means clustering algorithm as shown in Fig.1 which is a framework for the quality evaluation.

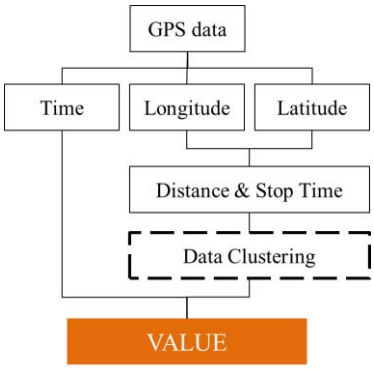

Figure 1. Framework for Quality Evaluation

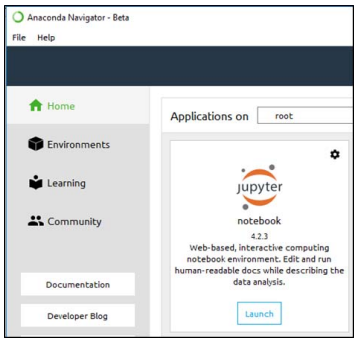

Figure 2. Anaconda startup screen

In this work, we use Python to implement K-means clustering algorithm. As a Python environment, we used Jupyter notebooks to interact with Python and the scientific Python stack. In order to call Jupyter from Windows software, "Anaconda (Figure 2. Anaconda startup screen)" was installed on a personal computer.Using such environments made it easy to run Python programs. The output result is displayed in the "Out" box by simply entering the program source in the "In" box and executing it (Fig 2. Jupyter (python) startup screen).

IV. Data Collection

The data were collected at Universiti Kebangsaan Malaysia (UKM) campus, which is included bus operation schedule, bus operation routes and bus operation frequency. The buses run throughout 9 square/km with 3 different routes and the capacity of the students on campus is about 11 thousand. Each bus route has been based as follows;

- Zone 2 (10 km),
- Zone 3U (11.5 km), and
- Zone 6 (11 km)

Operation schedule of buses on the three routes is as follows (Table I). "Number" stated in the table is the total number of buses running during the day, for three different area zones. Here below is buses operation plans.

1) *Bus service through the week:* All three routes of buses operate on weekdays. Only Zone 6 operates on public holiday and weekends.

2) *Bus service through the day:* 7:30 - 23:00, at 30 minute intervals (excluding 13:30 and 19:30)
3) *Zone 2:* two buses operate at the same time, for every 30 minutes, as per time line in (2).
4) *Zone 3U:* three buses operate at the same time, for every 30 minutes, as time line in (2).
5) *Zone 6:* one bus operates at the same time, for every 30 minutes, as per time line in (2).

TABLE I. NUMBER OF BUS

Type	Number	7:30-13:00	14:00-19:00	20:00-23:00
Zone 2	5	2	2	2
Zone 3U	8	3	3	2
Zone 6	3	1	1	1

The GPS data were collected based on the three (3) different elements as mentioned in previous sections in the area of UKM campus. It was collected using an application using Sports Tracker. The GPS data were collected as the following scheme and details.

- Bus routes: Zon 2, Zon 3U and Zon 6
- Record date and time: 8:00 to 22:00, at every hour trip, 15 trips per bus
- Frequency of data recording: 1 times /sec
- GPS data count: about 78,000 logs

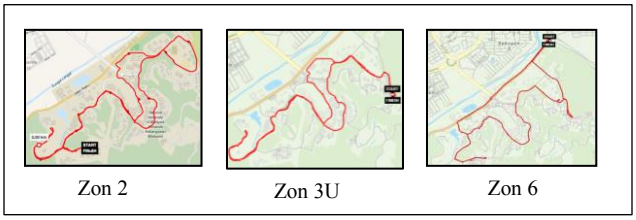

Figure 3. Types of Bus Route

The route for each zone of the buses operates is shown in Fig 3. There are a total about 28 bus stops along the route. However, when there are no passengers getting on or off, the bus will not stop and continue moving to the next bus stop. Sometimes, there are overlapping or sharing of the bus stops among the three routes where the buses operate. The recorded GPS data was in CSV format.

A Sports-Tracker is a GPS log application for smartphone that is used to get GPS data at 1 second intervals. The dataset obtained is shown in Fig. 4 below. From the GPS data and time stamps data, the tool used provides other data such as speed and distance, as shown in graph of Figure 4 plotted by the tool.

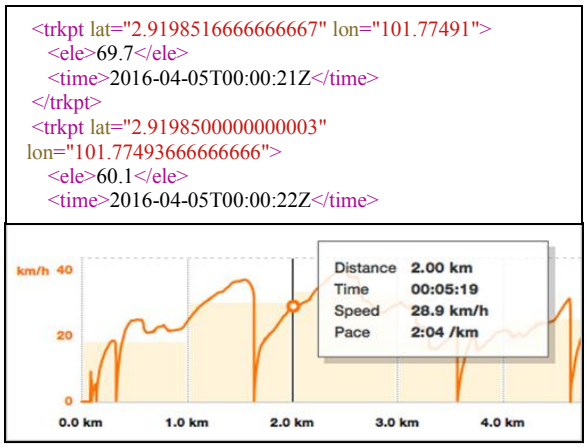

Figure 4. CSV dataset and Line graph

V. DATA PROCESSING AND ANALYSIS

The primary data set contains bus GPS location coordinates and the time stamps (one second intervals). In the following step, the data available were extended by calculating the number of stops made by bus through each trip. The duration of each stop is derived and added to the dataset. Table II shows a real sample of the number of stops recorded and the average of the total stops time. As described in the previous section has been set to get the GPS data every one second, and hence the distance moved was possible to measure. Calculation of the moving distance has adopted the generally calculation method, using the law of cosines (Hidetoshi, 2015).

TABLE II. GPS LOGS , SPEED AND BUS STATUS

Time	Distance	km/h	Stop
0:11:16	0.000657731	2.4	-
0:11:17	0.000519982	1.9	-
0:11:18	0.000413813	1.5	-
0:11:19	0.001078258	3.9	-
0:11:20	0	0.0	Stop
0:11:21	0	0.0	Stop
0:11:22	0	0.0	Stop
0:11:23	0	0.0	Stop
0:11:24	0	0.0	Stop
0:11:25	0	0.0	Stop
0:11:26	0	0.0	Stop
0:11:27	0	0.0	Stop
0:11:28	0.001905813	6.9	-
0:11:29	0.001189536	4.3	-
0:11:30	0.002112177	7.6	-

As it can be seen at Table II, a zero distance over sometime (8 seconds or above) is depicting a brief stop, likely a get off/on of passengers. Bus users getting on or off during this time. By summing up the following data from the data log; "Total Wait Time", "Count of Stops", and then "Average Wait Time" can be calculated for each time slot. In Table III, features are shown for each bus service in each time slot (eg, 8a.m., 14p.m., 20p.m.). For example, one can easily guess that there are many passengers on the Zone 2 bus at 8 o'clock and that at Zone 3U there are many passengers at 14 o'clock.

TABLE III. WAITING TIME

BUS		8:00	14:00	20:00
2	Total Wait Time	0:05:55	0:01:40	0:02:05
	Count of Stop	16	6	9
	Average Wait Time	0:00:22	0:00:17	0:00:14
3U	Total Wait Time	0:00:51	0:07:10	0:04:32
	Count of Stop	3	18	16
	Average Wait Time	0:00:17	0:00:24	0:00:17
6	Total Wait Time	0:01:14	0:02:32	0:01:07
	Count of Stop	3	8	3
	Average Wait Time	0:00:25	0:00:19	0:00:22

VI. ANALYSIS AND VISUALIZATION

Based on the proposed framework, the three (3) attributes such as time, volume and quality are described as follows:

- Time (x): at every hour from 8:00 to 22:00
- Volume (y): Number of stops during every hour trip
- Quality (z): duration of stops for every trip

For "Time" and "Volume" we used the data obtained from the GPS data. For the "Quality" attribute, the "total stop time" is calculated and could be used to draw assumptions, such as the number of passengers using the bus for the particular trip.

There are 28 bus stops throughout the three area zones. For each cycle, the buses may stop at 15 bus stops or more especially during peak hour time slots. For example, in Table IV, the number of stops at time 8a.m. and 10a.m, the number of stops are 16 times for zone 2. However, there are times when the bus stops less than 5 times. When the total stop time (Z-quality) is longer i.e 10a.m. for zone 2, it is showing high possibility that there are many passengers. Table IV shows the data accumulated in different colour is to make the size of the data easy to understand.

A. K-means Clustering

K-means is a widely used unsupervised clustering algorithm, suggested (MacQueen, (1967); Pang-Ning Tan, 2014). The goal of clustering is to produce none-overlapping groups of data set with a high degree of similarity within each group, with the number of groups or clusters by the variable K. Such grouping enable bring more insights to the interpretation of data. Another goal of K-means clustering is to reduce the complexity of data set (A. K. Jain, 1999). The steps for K-means algorithm are as follow:

1. Select K points as initial centroid.
2. Assign all points of data to the closest centroid, forming K clusters
3. Compute again centroid of each cluster
4. Repeat 2 and 3 till centroids do not change

In this experiment, the operation level of K-means is classified as "High", "Middle" and "Low", as shown in Fig. 5. The result of clustering is depicted in Fig. 6 where the optimal parameter setting is based on the middle operation level. The result shows that data with high operation is classified in group 2 and data with low operation is classified in group 0.

TABLE IV. THREE IDENTIFIED ATTRIBUTES FOR EACH ZONE

Zon 2			Zon 3U			Zon 6		
X:time	Y:volume	Z:quality	X:time	Y:volume	Z:quality	X:time	Y:volume	Z:quality
TIME	STOP	WAIT	TIME	STOP	WAIT	TIME	STOP	WAIT
8:00	16	0:05:55	8:00	3	0:00:51	8:00	3	0:01:14
9:00	4	0:00:59	9:00	15	0:05:35	9:00	5	0:01:55
10:00	16	0:06:26	10:00	17	0:06:40	10:00	4	0:00:59
11:00	12	0:04:40	11:00	3	0:00:50	11:00	12	0:03:59
12:00	11	0:04:38	12:00	5	0:01:43	12:00	6	0:01:40
13:00	7	0:01:13	13:00	5	0:01:40	13:00	10	0:02:19
14:00	6	0:01:40	14:00	18	0:07:10	14:00	8	0:02:32
15:00	6	0:01:12	15:00	13	0:04:34	15:00	6	0:01:39
16:00	10	0:02:56	16:00	11	0:03:16	16:00	8	0:02:21
17:00	5	0:01:00	17:00	10	0:04:54	17:00	8	0:02:20
18:00	6	0:01:38	18:00	7	0:02:18	18:00	8	0:03:23
19:00	5	0:01:10	19:00	6	0:01:12	19:00	6	0:01:15
20:00	9	0:02:05	20:00	16	0:04:32	20:00	3	0:01:07
21:00	4	0:00:53	21:00	7	0:01:46	21:00	5	0:01:17
22:00	3	0:00:52	22:00	12	0:03:33	22:00	8	0:02:55

```
# K-means Clustering
kmeans_model =
KMeans(n_clusters=3,
random_state=10).fit(features_zon2)

# Get a label
labels = kmeans_model.labels_

# Display the result
for label, features_zon2 in zip(labels,
features_zon2):
   print(label, features_zon2)
```

Figure 5. K-means Clustering

```
# Result of Clustering
2 [ 16 355]
1 [ 4 59]
2 [ 16 386]
2 [ 12 280]
2 [ 11 278]
1 [ 7 73]
1 [  6 100]
1 [ 6 72]
0 [ 10 176]
1 [ 5 60]
1 [ 6 98]
1 [ 5 70]
0 [  9 125]
1 [ 4 53]
```

Figure 6. Clustering result

All bus data were classified according to the above procedure where the numbers were assigned to Clusters are set randomly. The issue would be the size of each cluster. The result of classification is shown in Table V.

Zon 2: (1 < 0 < 2), Zon 3U: (2 < 0 < 1), Zon 6: (1 < 0 < 2)

TABLE V. SAMPLE OF CLUSTERING RESULT FOR THREE ZONES

Time	# Zon 2	# Zon 3U	# Zon 6
8:00	2 [16 355]	2 [3 51]	1 [3 74]
9:00	1 [4 59]	1 [15 335]	0 [5 115]
10:00	2 [16 386]	1 [17 400]	1 [4 59]
11:00	2 [12 280]	2 [3 50]	2 [12 239]
12:00	2 [11 278]	2 [5 103]	1 [6 100]
13:00	1 [7 73]	2 [5 100]	0 [10 139]
14:00	1 [6 100]	1 [18 430]	0 [8 152]
15:00	1 [6 72]	0 [13 274]	1 [6 99]
16:00	0 [10 176]	0 [11 196]	0 [8 141]
17:00	1 [5 60]	0 [10 294]	0 [8 140]
18:00	1 [6 98]	2 [7 138]	2 [8 203]
19:00	1 [5 70]	2 [6 72]	1 [6 75]
20:00	0 [9 125]	0 [16 272]	1 [3 67]
21:00	1 [4 53]	2 [7 106]	1 [5 77]
22:00	1 [3 52]	0 [12 213]	0 [8 175]

Whereas, Table VI shows the data that have been sorted by the size of cluster.

TABLE VI. SORTED RESULT OF CLUSTERING

# Zon 2	# Zon 3U	# Zon 6
2 [16 355]	1 [15 335]	2 [12 239]
2 [16 386]	1 [17 400]	2 [8 203]
2 [12 280]	1 [18 430]	0 [5 115]
2 [11 278]	0 [13 274]	0 [10 139]
0 [10 176]	0 [11 196]	0 [8 152]
0 [9 125]	0 [10 294]	0 [8 141]
1 [4 59]	0 [16 272]	0 [8 140]
1 [7 73]	0 [12 213]	0 [8 175]
1 [6 100]	2 [16 51]	1 [3 74]
1 [6 72]	2 [3 50]	1 [4 59]
1 [5 60]	2 [5 103]	1 [6 100]
1 [6 98]	2 [5 100]	1 [6 99]
1 [5 70]	2 [7 138]	1 [6 75]
1 [4 53]	2 [6 72]	1 [3 67]
1 [3 52]	2 [7 106]	1 [5 77]

Based on the results shown in Table VI, we were able to obtain further insights of the operation details in each of the three area zones. It can be understood that there are several time slots in which Zon 2 runs less, and Zon 3U and Zon 6 may be used more frequently than Zon 2. Also, as appears from the table above, that there are several situations in which the use time of each bus is very limited.

B. Visualisation

The 3-attributes approach is extended to the visual display and analysis. Graphs in Fig. 7 shows a multi-dimensional display of the 3-attributes (x, y, z) for zone 2. According to the bus rules of procedure, the bus begins the trip from the Start station, and would stop at intermediate bus stops only in case there is a passenger to get on or off. Therefore, in these cases the bus would continue moving to other intermediate bus stops without stop by.

The cycle of each trip on the route to the End station usually takes between 30 to 40 minutes. Time of departure are adjusted and defined at the Start and End stations. Due to that fact, we use the stop time at each stop station as an indicator for the related number of passengers using the bus. The size and color for each dot in Fig. 7 depicted to the length of stop point in seconds, hence the number of passengers using the bus at the specific bus trip.

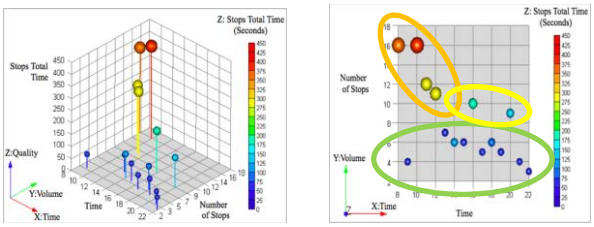

Figure 7. Bus Operation of 1day (Zon 2: x, y, z)

VII. RESULT AND DISCUSSION

Our approach used three distinct attributes from our dataset, namely, bus hourly time, number of stops during each trip, and time in seconds for stops during each trip in the day. Those three elements of data are used to express in a simple visual way of the bus operation for each route. The analysis and information that are being visible at graphs, can give a simple and expressive vision of the daily operation station for buses. As with this case operation, there are 1, 2 or 3 vehicles available and run at each trip at the three routes. Therefore, the visual evaluation as the one presented in this paper can be used to revise the number of vehicles deployed at each time every day and makes a proper decision. The application of the approach presented in this paper, provides two options of machine-friendly numerical values, along with human-friendly visual display.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we have presented a new approach for a quick, simple and handy approach for analyses data as decision support both for bus service provider as well as

for customers or users. A typical handy mobile phone GPS tool was used to collect data. Three attributes of data, resembling Time, Volume of iterated instances, and Quality data that expresses the most interest are used to have a simple but a concrete evaluation of information or services for review, planning, or daily decision-making. Application of the approach to the bus operation in university campus was done by demonstrating the process of application starting from a selection of simple data collection method, refining and expressing data obtained to specify Time, Volume and Quality attribute data. Following, would be the visualisation of the point of interest such as bus operation. Big data and AI will be the center of the future for many data sciences. However, that would not be available for all stakeholders to apply, and integrate in business or service planning and operation, due to costs and skills required to come out with quick insights. This research aimed at a handy, simple, and affordable solution for big data analysis. Our future work will be directed to wider transportation networks, develop an online application for analysis and visualisation of operation of the bus, and provision of service for end users. Other work will be towards application of the framework and approach proposed for a wider range of different areas of the data set, and finally, develop more standard and generic framework that would be useful in simplifying big data analysis.

ACKNOWLEDGEMENT

This research has been supported by the Fundamental Research Grant Scheme funded by the Ministry of Higher Education (MoHE), Malaysia under Project Code FRGS/1/2017/ICT04/UKM/02/8. It is a collaborative research between the Institut of Visual Informatics (IVI), Universiti Kebangsaan Malaysia and Meio University, Japan.

REFERENCES

[1] Mohamed Rehan Karim. Achieve competition and share with each well-balanced transportation agency: 3rd International Symposium on Intercity Transport System in Asian Countries. *Institute for Transport Policy Studies*, Vol.15 No.2; 2012, pp.113-114 (In Japanese).

[2] Natsumi Abe. A Study on the City for the Knowledge Society: In the case of Cyberjaya, Malaysia. Summaries of technical papers of annual meeting, 2013, pp.1175-1176 (In Japanese)

[3] Suwardo, Madzlan Napiah, Ibrahim Kamaruddin. ARIMA MODELS FOR BUS TRAVEL TIME PREDICTION. The Journal of the Institution of Engineers, Malaysia, vol. 71(2), 2010, pp. 49-58

[4] Mohammad Hesam Hafezi, Amiruddin Ismail. Interaction between Bus Stops Location and Traffic on Bus Operation. Applied Mechanics and Materials, Vols. 97-98, 2011, pp. 1185-1188

[5] Tsutomu Yabe. A Study on Planning Process for Bus Rapid Transit System. Yokohama National University, March 24, PhD. (Engineering), 2006, B No. 259(In Japanese)

[6] Gavin Kemp, Genoveva Vargas-Solar, Catarina Ferreira da Silva, Parisa Ghodous, Christine Collet, Pedropablo Lopez. Towards Cloud big data services for intelligent transport systems. Concurrent engineering, hal-01213302, Delft, Netherlands, 2015.

[7] Jeanne W. RossCynthia M. BeathAnne Quaadgras. You May Not Need Big Data After All. Harvard Business Review, 2013.

[8] Munzilah Md. Rohani, Devapriya Chitral Wijeyesekera, AhmadTarmizi Abd.Karim. Bus Operation, Quality Service and The Role of Bus Provider and Driver. Procedia engineering, Vol. 53, 2013, pp.167-178.

[9] Jayakrishna PATNAIK, Athanassios BLADIKAS, Using Data Mining Techniques on APC Data to Develop Effective Bus Scheduling Plans. SYSTEMICS, CYBERNETICS AND INFORMATICS VOLUME 4 - NUMBER1; 2006, pp.86-90

[10] Peter G. Furth. Data Analysis for Bus Planning and Monitoring, A Synthesis of Transit Practice, National Academy Press, Washington, D.C., 2000.

[11] Andrei Iu. Bejan, Richard J. Gibbens, David Evans, Alastair R. Beresford, Jean Bacon. Statistical Modelling and Analysis of Sparse Bus Probe Data in Urban Areas. 13th International IEEE Conference on Intelligent Transportation Systems, Madeira Island, Portugal, 201,pp.19-22.

[12] Leon Stenneth, Philip S. Yu. Monitoring and mining GPS traces in transit space. Proceedings of the 2013 SIAM International Conference on Data Mining, 2013, pp. 359-368.

[13] S. L. Bangare, A. D. Kadam, P. S. Bangare, P. V. Katariya, C. A. Khot, N. R. Kankure. Solutions Concerning Information Systems for Real Time Bus Arrival. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, 2013, pp. 316-319

[14] Gang Zeng. Application of Big Data in Intelligent Traffic System. IOSR Journal of Computer Engineering (IOSR-JCE), Volume 17, Issue 1, Ver. VI, 2015, pp 01-04.

[15] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and, C. Shahabi. Big Data and Its Technical Challenges, Communications of the ACM, Vol. 57, No. 7., 2014.

[16] Hidetoshi Miura. Three distance calculation method using the latitude and longitude", Unauthorized reproduction of this article is prohibited (In Japanese)., 2015.

[17] MacQueen, J. Some methods for classification and analysis of multivariate observations, Proceedings of the Fifth Berkeley Symposium On Mathematical Statistics and Probabilities, 1, 1967, pp. 281-296.

[18] Pang-Ning Tan, Michael Steinbach, Vipin Kumar. (2014). Introduction to data Minning, Pearson Eduction, 2014.

[19] A.K. Jain, M. N. Murty, P. J. Flynn. Data Clustering: A Review, ACM Computing Surveys , Volume 31 (3), 1999, pp. 264–323