APPLICABILITY OF CROWD SOURCING TO DETERMINE THE BEST TRANSPORTATION METHOD BY ANALYSING USER MOBILITY

J.M.D. Senanayake[1] and W.M.J.I. Wijayanayake[2]

[1]Department of Industrial Management, University of Kelaniya, Sri Lanka
[2]Department of Industrial Management, University of Kelaniya, Sri Lanka

ABSTRACT

Traffic is one of the most significant problem in Sri Lanka. Valuable time can be saved if there is a proper way to predict the traffic and recommend the best route considering the time factor and the people's satisfaction on various transportation methods. Therefore, in this research using location awareness applications installed in mobile devices, data related to user mobility were collected by using crowdsourcing techniques and studied. Based on these observations an algorithm has been developed to overcome the problem. By using this, the best transportation method can be predicted as the results of the research. Therefore, people can choose what will be the best time slots & transportation methods when planning journeys. Throughout this research it has been proven that for the Sri Lankan context, the data mining concepts together with crowdsourcing can be applied to determine the best transportation method.

KEYWORDS

Big Data, Crowd sourcing, Data mining, GPS, IoT

1. INTRODUCTION

With the technological enhancements related to Internet, Wireless Communication, Big Data Analytics, Sensors Data, Machine Learning; a new paradigm is enabled for processing large amount of data which are collected from various sources. Internet of Things (IoT) is one of the great source that can be used to get a huge stream of data. IoT is a platform and an evolving technology that allows anything which connected to Internet, to process information, communicate data, analyse context collaboratively and in the service or individuals, organizations and businesses.

In the past decades, both coarse and fine-grained sensor data had been used to perform location-driven activity inference. On one hand, a strand of related work attempt to recognize individual activity using the data collected by a cluster of wearable sensors. Although the recognition performance is relatively high, the human efforts on carrying many extra sensors are still open challenges. In recent years, GPS phone or GPS enabled PDA become an essential in people's daily lives. With such devices it has become very easy to trace people's outdoor mobility using location-based applications.

Modelling big data is a current trend and combining that with the Internet of Things & crowdsourcing is an interesting area for a research work. In this research the data related to user locations, were collected using the devices such as mobile phones etc. which were connected to the Internet, were mined using data mining techniques and came up with an algorithm to model & analyse those big data to identify mobility pattern, to determine best routes, to find the best transportation method considering traffics & to find transportation method satisfaction etc.

DOI: 10.5121/ijdkp.2018.8503

2. RELATED WORK

With the rapid growth in the technological industry, mobile devices become an essential item of people's lives due to the fact that these devices provide a lot of features which help people to do their routine activities in much easier way. Therefore, society expects more from these devices which leads device vendors & application developers to do more R&D and enhance the features. Due to this, mobile devices which were having the functionality of just making a call became devices which can perform "smart activities". They have become more powerful with the integration of chip sets such as global positioning systems (GPS) to measure geospatial location and accelerometers that can measure a devices orientation.

Even though the popularity of the GPS is high, in some researches it has claimed that using GPS as a method to collect user mobility data and geographical data is not much accurate since GPS data may contain long gaps & poor user time coverage because signal establishment with the satellites are cut off in indoors or due to the impenetrable covers where people stay most of their time. And the average time associated with GPS coverage was a mere 4.5%, whereas GSM and 802.11 coverage were 99.6% and 94.5%, respectively [1].

The spatial accuracy coupled with the high accuracy clocks of GPS satellites allows for a great representation of a user's mobility. There are 24 satellites (plus some additional satellites) orbiting the Earth in six different planes. Each of these planes are inclined 55◦ from the Earth's equatorial plane. These satellites are positioned in their respective planes in way that at least four of them are above the horizon every time from almost any place on the Earth. As long as GPS devices are having at least 4 partial view of the sky, they are not having any problems of receiving signals [2]. By assuming that there are no obstructions, to get an accurate fix by GPS receivers at any given time they are nearly always guaranteed to be in view of the minimum number of 4 satellites. If enough satellites are in view, an accuracy within two meters can be achieved (5-10 meters is a realistic expectation [3]).

The spatial data collected using GPS are need to be processed before using. Processing is defined as repairing or putting through a prescribed procedure which contains the steps of filtering, smoothing and interpolation. The main reason for processing this collected data is to replace the impossible task of visually inspecting the collection. When processing each step performs an essential task which determines unfavourable attributes and either identifies or removes them [4].
There are already implemented ways for controlling & managing traffic such as safety cameras and other existing traffic management methods. But they are not good enough for every situation & every location due to the complexity of traffic networks, traffic speed and the huge number of traffic participants [5]. Therefore, the research findings [5] described a new traffic management solution based on the automatically individual control to any traffic user anywhere and anytime. The system can establish traffic management because of this traffic management algorithm which has the following principle. "The central traffic management unit get to know about the location, speed and condition for every single registered vehicle by decoding and analysing information about itself which were periodically sent".

A simple yet very effective method that can capture traffic states in complex urban areas has also been proposed [6]. In that, they applied their methodology to two different GPS trace data sets which were collected in the Ann Arbor in Michigan. They have found out that higher than 90% accuracy can be achieved if 10 or more traversal traces are collected on each road based on the results. In addition, traffic patterns turned out to be fairly consistent over time, which allowed the use of a larger history in classifying traffic conditions.

A technique to identify road traffic congestion levels from velocity of mobile sensors with high accuracy and consistent with motorists' judgments has been proposed in another research [7]. At the data collection stage they have used a GPS device and a webcam. An opinion survey has also been used in this research to rate the traffic congestion levels into three levels: light, heavy, and jam. The ratings and velocity were fed into a decision tree learning model and they successfully extracted vehicle movement patterns to feed into the learning model using a sliding windows technique. The accuracy of this model was 91.29%.

A stream computing approach was used in another research for a real-time Traffic Information Management methodology [8]. For this GPS data from some taxis and trucks were used to showcase some of their findings on traffic variability in the city of Stockholm. Traffic volume measurements by region, estimates of travel times between different points of the city, continuously updated speed and traffic flow measurements for all the different streets in a city, stochastic shortest path routes based on current traffic conditions, etc. were included in their customized analysis. In addition to that time-dependent travel time estimates have to be integrated into time dependent vehicle routing frameworks to benefit from telematics based data collection. For time dependent vehicle route planning framework, it has been discussed about data collection and the conversion from raw empirical traffic data into information models, an application example which compares several information models based on real traffic data regarding its benefits, the integration of information models into it. A data mining approach followed in this research provides time-dependent travel times in a memory efficient way without a significant reduction of the itineraries' reliability and robustness [9].

By analysing two data clustering algorithms: The K-Means Clustering, and the Fuzzy C-Means Clustering a methodology has been presented for detection of hot spots of traffic through analysis of GPS [10]. In this methodology a cluster centre can be selected once the clustering process stops. This will display the membership grades of all data points toward the selected cluster centre. It has been justified in this research that the fact of using clustering algorithm for the detection of the hot-spots, where each cluster represents the group of GPS data points having latitude and longitude as their coordinate and having very small distance between them. A formula has also been derived in order to calculate geodesic distance between a pair of latitude/ longitude points on the surface of the earth, using the WGS-84 (World Geodetic System -84) ellipsoidal which comprise of a reference ellipsoid, a standard coordinate system, altitude data and a geoid [10].

3. METHODOLOGY

As objectives of this research we need to experimentally determine the traffic predictions & designing an algorithm to find the best transportation method. Survey research approach is not in line with research objectives, as this study is for generating algorithms related to efficient transportation system in Sri Lanka. Therefore, Experimental research approach is more preferable in this case. As in the Figure 1 this research was sub divided into several activities: data collection, processing and segmentation, map matching, clustering, aggregation, and model training and transport method prediction.

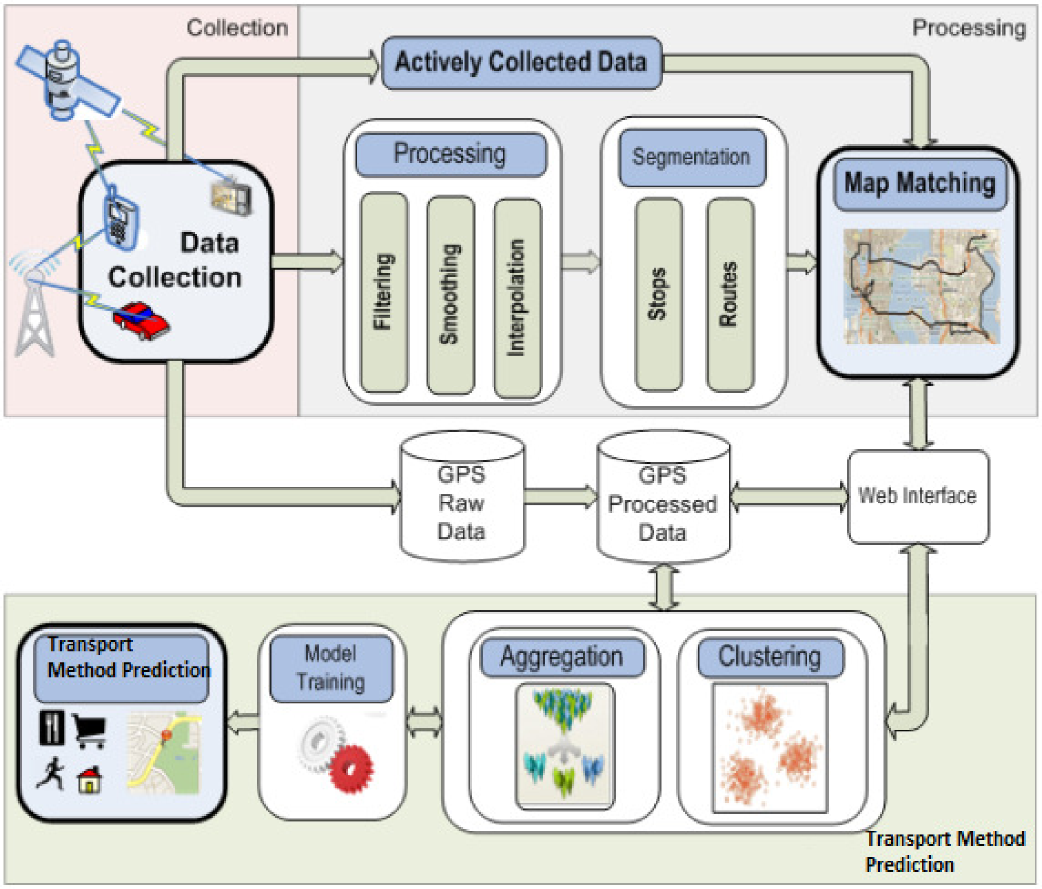


Figure 1: Research Design

3.1. Data Collection

Data Collection part was done by using an Android application named as “Best Method” which was installed on smartphones of the users who were willing to participate in location sharing & transport method satisfaction sharing project. Crowdsourcing techniques [11] had been used for the purpose of data collection. In this application the user had to choose the transportation method that the user was using and the satisfaction of current transportation method by giving a rating while the mobile phone was connected to Internet and the status of the GPS Service was turned to on. The collected data was uploaded to a cloud based storage and used in the later steps.

Individuals were asked to use this application,

- When they were travelling between Bambalapitiya and Pettah, Bambalapitiya and University of Colombo, Pettah and University of Colombo.
- In the period of 1st November 2017 to 30th November 2017 and in the time periods of 7 AM to 8.30 AM & 5 PM to 6.30PM.

The users had to select the transportation method which was currently using and the satisfaction of that transportation method. Users were asked to rate the transportation method satisfaction based on the traffic which was happening at the time.

The location of the device & the current date & time were picked automatically from the device. Apart from that to determine the GPS Signal strength, additional measurement was also checked how much satellites are connected to the device was. If the number of connected satellites was greater than or equal to four (which was the decision point of the signal strength as studied), then the GPS Strength will be taken as Strong or Otherwise Weak. After that the collected data were stored on a cloud storage.

3.2. PROCESSING & SEGMENTATION

Processing was done under 3 steps known as filtering, smoothing & interpolation and after that the segmentation was started.

If the moving speed is <1 mph and the no of GPS data in a consecutive time period is low that data will be filtered out programmatically [12]. There will be an exception when the transportation method of the filtering row is “Walking”. In that case the rule of speed <1 mph can’t be applied. Therefore, traffic prediction will not be applicable if the transportation method is walking. But that data will not be filtered out since it is used for the best transportation method selection process. The Extended Kalman filter [13] was used in order to smooth the GPS data set. In the event of signal loss by imputing missing data values between two logged points is the method that used to interpolate. This method is to insert values at 5-minute intervals (depending on the time gap). If two temporally adjacent points bounding a period of missing data were within 30-meters (the distance used to determine the two points were in the same or similar location), the missing GPS data point to the earlier point is assigned.

3.3. MAP MATCHING

In the context of this system, the map-matching module will be used as a refinement step to make the final adjustments to each GPS point ensuring they are usable and correct. Another purpose behind the map-matching module is to remove large portions of the collected data without jeopardizing the integrity of the data. This step performs the bulk of the logic behind all of the filtering. It turns a raw GPS trajectory into a reduced (by removal of unnecessary points) and adjusted (points are snapped to the road network) route. This is achieved by using a modified Douglass-Peucker algorithm [14].

3.4. TRANSPORT METHOD PREDICTION

After the GPS data is processed and verified, the remaining modules perform the tasks necessary to start training prediction models. The clustering module uses the K-Means algorithm which has been identified as the suitable way [15] to cluster the processed data based on the location and the transportation method.

The last two modules are used for model training and finally prediction/labelling. The training module currently utilizes the decision tree algorithm to predict the traffic.
Based on the results the best transportation was determined.

4. RESULTS/DISCUSSION

4.1. DATA ANALYSIS

At the end of the period the cloud storage contains around two hundred & sixty thousand records which contains individuals’ data. The collected data set contained individuals’ spatial data (latitude & longitude along with the time stamp) along with other necessary data (device ID, transportation method, satisfaction and the GPS Strength).

As the first step a PHP script was run to filter out the records which has weak GPS strength. The SQL query to perform that as follows.

Select * from user_data where gps_stregth='strong';

After that another table was filled with that selected data set which contains around two hundred & fifty thousand records. In these records there wasn't field to say where the user is heading. For an example the user may travel from Bambalapitiya to Kollupitiya or Kollupitiya to Bambalapitiya. To determine direction another PHP script has been used based on the consecutive GPS data points of individual users.

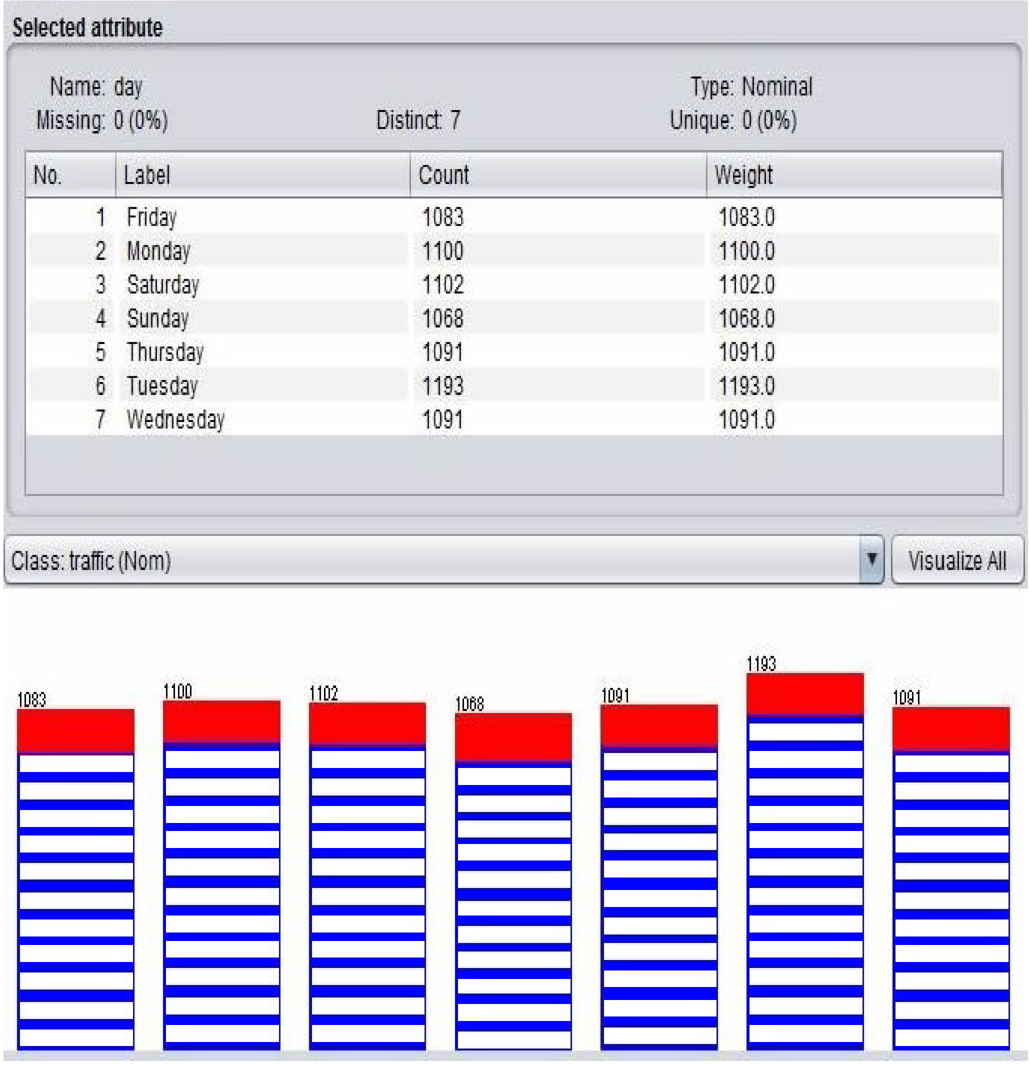
The next step was to identify the location based on the latitude & the longitude of the records. Since the number of records is also high and there is no point of analyzing each & every latitude & longitude [16]. The GPS locations were clustered using K-Means clustering algorithm to derive the spatial data to more meaningful stage. The K-Means algorithm implemented using PHP is used for this purpose. After the new derived table was generated using the clustered data, next step is using the Google Geocoding API. Since the data set contains the geographic coordinates, reverse geocoding was used to convert geographic coordinates into a human-readable address. The Google Maps Geocoding API has a usage limit of 2500 request per day for a standard user. Since this research is an educational one but not a commercialized one the standard usage is the used method. Therefore, when determining the exact location based on the number of data it will take around few days to complete. (For the analysis it took 5 days since after clustering there were around 120000+ records).

That was performed using a PHP script together with a Javascript. For this the dataset was converted into a CSV format and the algorithms were performed next and the stored the values again in the previous MySQL table.

Then based on the average ratings of the satisfaction of the transportation method and the moving speed, the traffic status was generated. And relevant MySQL table was also updated with those results.

Figure 2 visualize the status of the attributes & records in this aggregate data set which is using the "Best Transportation Method" application. In the following visualization ☰ represents that there is no traffic and 🟥 represents that there is traffic based on the prediction.

Now the database table named as "data_aggregation_traffic" is stable for retrieving a meaningful & useful output. This new table is used together with the end user application (www.bestmethod.esy.es). In this application the user has to select the parameters accordingly and based on the research results the best transportation methods will be recommended.



Selected attribute

Name: day Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 0 (0%)

No.	Label	Count	Weight
1	Friday	1083	1083.0
2	Monday	1100	1100.0
3	Saturday	1102	1102.0
4	Sunday	1068	1068.0
5	Thursday	1091	1091.0
6	Tuesday	1193	1193.0
7	Wednesday	1091	1091.0

Class: traffic (Nom)

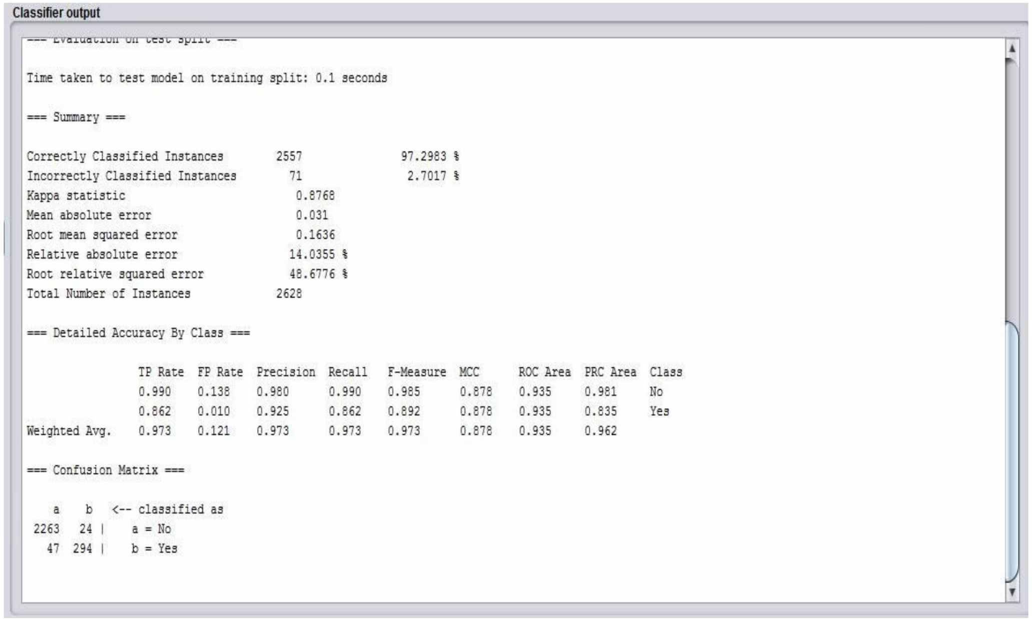
Figure 2: Visualization of the daily traffic prediction in a week

4.2. VALIDATION

After the period of time (1/11/2017 – 30/11/2017). Another android mobile application was distributed to the same set of users, only for collecting the data to validate the research. In this mobile application the user just has to select the Transportation method together with the current location and whether there is a traffic or not.

The data were collected and stored for a period of one week (1/12/2017- 7/12/2017) separately from the above research data and using Weka libraries two data sets were analyzed using decision tree classification algorithm to validate the results. The original research dataset was used as the training data set and the newly collected dataset was used as the testing dataset (Figure 3).

Based on the Weka analysis, it has given the correctly classified instances with a percentage of above 97 (97.29%). This has proven that the result of the research can properly be validated.



```
=== Evaluation on test split ===

Time taken to test model on training split: 0.1 seconds

=== Summary ===

Correctly Classified Instances        2557               97.2983 %
Incorrectly Classified Instances        71                2.7017 %
Kappa statistic                          0.8768
Mean absolute error                      0.031
Root mean squared error                  0.1636
Relative absolute error                 14.0355 %
Root relative squared error             48.6776 %
Total Number of Instances             2628

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.990    0.138    0.980      0.990    0.985      0.878    0.935     0.981     No
                 0.862    0.010    0.925      0.862    0.892      0.878    0.935     0.835     Yes
Weighted Avg.    0.973    0.121    0.973      0.973    0.973      0.878    0.935     0.962

=== Confusion Matrix ===

    a    b   <-- classified as
 2263   24 |    a = No
   47  294 |    b = Yes
```

Figure 3: Research validation using training data set & testing data set

4. RESULTS/DISCUSSION

The main objectives of this research were Applying data mining and big data analytics techniques for a database which contains IoT data related to user movements, identify mobility patterns of the users, and identify traffic times of the interested areas based on the big data analytics models and applying data mining techniques to determine most suitable transportation method considering all the constraints such as traffic, crowd etc. based on the previous experiences of people.

The research was done considering the mobility patterns of the people who were travelling in the Colombo area. And after the validation process, it has proven that the predicted traffic and the recommended transportation method is the most suitable for the given parameters. Since the predicted output is valid, by increasing the data collection ranges and the audience this research can be further extended into more meaningful stage.

There are mainly two parties who will be benefited from this research. They are the general public & the government. The general public will be benefited because they can use the "Best Transport Method" application which is the final outcome of this research to plan their journeys before begin them. The government will also be benefited when it comes to the planning smart cities based on eliminating the traffic to a certain extent.

Using the same original data set we can expand the research areas into few other major areas like determining the travel patterns of the people and identify the places where the people usually travels and the places which the traffic can be occurred regular and take necessary smart action based on the outcome.

Based on the research findings, we can apply the same methodology to predict the best transportation method in all the areas in Sri Lanka by extending the data set of the collection stage and based on a thorough analysis, since it has been proven that to determine the best transportation method in Sri Lankan context, data mining concepts together with crowdsourcing can be applied throughout this research.

REFERENCES

[1] A. Lamarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert , P. Powledge, G. Borriello and B. Schilit, “Place lab: Device Positioning Using Radio Beacons in the Wild,” in Proceedings of the Third International Conference on Pervasive Computing, May 2005, pp. 116–133.

[2] N. Ashby, “Relativity in the Global Positioning System,” Living Reviews in Relativity, vol. 6, no. 1, 2003.

[3] J. Wolf, “Applications of New Technologies in Travel Surveys,” in Travel Survey Methods, P. Stopher and C. Stecher, Eds., 2006, pp. 531 - 544.

[4] J. Jun, “Smoothing methods designed to minimize the impact of gps random error on travel distance, speed, and acceleration profile estimates,” in Ph.D. thesis, Department of Civil Engineering, Clemson University, 2005.

[5] A. Kardashyan and A. Kardashyan, “New Concept of the Urban and Inter-Urban Traffic,” in Proceedings of the World Congress on Engineering and Computer Science, San Francisco, USA, 2011, pp. 833-837.

[6] J. Yoon, M. Liu and B. Noble, “Surface Street Traffic Estimation,” in Proceedings of the 5th international conference on Mobile systems, applications and services MobiSys '07, San Juan, Puerto Rico, USA, 2007.

[7] T. Thianniwet, S. Phosaard, member, IAENG and W. Pattara-Atikom, “Classification of Road Traffic Congestion Levels from GPS Data using a Decision Tree,” in Proceedings of the World Congress on Engineering, London, U.K., 2009, Vol I.

[8] A. Biem, E. Bouillet, H. Feng, A. Ranganathan, A. Riabov and O. Verscheure, “Real-Time Traffic Information Management using Stream Computing,” IEEE Data Engineering Bulletin, vol. 33, no. 2, pp. 64-68, 2010.

[9] J. F. Ehmke and D. C. Mattfeld, “Data allocation and application for time-dependent vehicle routing in city logistics,” European Transport \ Trasporti Europei, no. 46, pp. 24-35, 2010.

[10] J. Tripathi, “Algorithm for Detection of Hot Spots of Traffic through Analysis of GPS Data,” in MSc. Thesis, Computer Science and Engineering Department, Thapar University, Patiala., 2010.

[11] E. Estelles Arolas and F. González-Ladrón-De-Guevara, “Towards an integrated crowdsourcing definition,” Journal of Information Science, no. 32(2), pp. 189-200, 2012.

[12] J. Froehlich and J. Krumm, “Route Prediction from Trip Observations,” SAE Technical Paper, no. 2008-01-0201, 2008.

[13] Manon Kok, Jeroen D. Hol and Thomas B. Schon, “Using Inertial Sensors for Position and Orientation Estimation,” Foundations and Trends in Signal Processing, vol. 11, no. 1-2, pp. 1-153, 2017.

[14] R. Ivanov, “Real-time GPS track simplification algorithm for outdoor navigation of visually impaired,” Journal of Network and Computer Applications, vol. 35, no. 5, pp. 1559-1567, 2012.

[15] M. Bhatia and D. Khurana, “Experimental study of Data clustering using k-Means and modified algorithms,” International Journal of Data Mining & Knowledge Management Process, vol. 3, no. 3, pp. 17-30, 2013.

[16] B. Varghese, A. Unnikrishnan and K. Jacob, “Spatial Clustering Algorithms- An Overview,” Asian Journal of Computer Science And Information Technology, vol. 3, no. 1, pp. 1-8, 2013.

AUTHORS

J.M.D. Senanayake holds a BSc. (Special) Degree in Management & Information Technology – Major in IT with a 1st Class Honours from University of Kelaniya, Sri Lanka. He worked as a Software Engineer at DirectFN | Mubasher (Pvt.) Ltd. and currently working as a Software Engineer attached to the Software Development & R&D Division at National Development Bank, Sri Lanka. His research interests are Software Engineering, Advanced programming languages, High performance real time software systems, Web Analytics & Web Engineering, Data Mining & Advanced Databases and Information Security

Prof. W.M.J.I. Wijayanayake received a PhD in Management Information Systems from Tokyo Institute of Technology Japan in 2001. He holds a Bachelor’s degree in Industrial Management from the University of Kelaniya, Sri Lanka and Master’s degree in Industrial Engineering and Management from Tokyo Institute of Technology. He is currently a Professor at the Department of Industrial Management, University of Kelaniya, Sri Lanka. His research interests are Information System Engineering, Data Engineering, Software Engineering, Business Intelligence, and Knowledge Management.