Communication

Travel Time Prediction Based on Data Feature Selection and Data Clustering Methods

Chi-Hua Chen [1,*]

[1] Department of Information Management, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan; chihua0826@nkfust.edu.tw

* Correspondence: chihua0826@nkfust.edu.tw; Tel.: +886-7-6011000 ext. 4126

Abstract: In recent years, governments applied intelligent transportation system (ITS) technique to provide several convenience services (e.g., garbage truck app) for residents. This study proposes a garbage truck fleet management system (GTFMS) and data feature selection and data clustering methods for travel time prediction. A GTFMS includes mobile devices (MD), on-board units, fleet management server, and data analysis server (DAS). When user uses MD to request the arrival time of garbage truck, DAS can perform the procedure of data feature selection and data clustering methods to analyses travel time of garbage truck. The proposed methods can cluster the records of travel time and reduce variation for the improvement of travel time prediction. After predicting travel time and arrival time, the predicted information can be sent to user's MD. In experimental environment, the results showed that the accuracies of previous method and proposed method are 16.73% and 85.97%, respectively. Therefore, the proposed data feature selection and data clustering methods can be used to predict stop-to-stop travel time of garbage truck.

Keywords: data feature selection; data clustering; travel time prediction

1. Introduction

Intelligent transportation system (ITS) has been more and more popular in recent years. Government applied ITS technique to provide several convenience services (e.g., garbage truck app (GTA) [1], bus app, public bicycle system, mass rapid transit system, and railway app) for residents. For instance, GTA can provide the location information of garbage truck and arrival time of each stop for reducing citizen's waiting time. For arrival time and stop-to-stop travel time prediction, some approaches which included statistical methods, linear regression methods, and neural networks were proposed and evaluated [1-3]. However, the computation power and time are needed by neural networks. Although statistical methods and linear regression methods can provide the predicted information quickly, the accuracy of predicted information may be lower when a larger variation exists in historical records.

Therefore, this study proposes a garbage truck fleet management system (GTFMS) and data feature selection and data clustering methods for travel time prediction. A GTFMS includes mobile devices (MD), on-board units (OBU), fleet management server (FMS), and data analysis server (DAS). OBU can detect the location and stop-to-stop travel time and send these records to FMS via cellular networks. FMS can store these records from OBU into database. When user uses MD to request the arrival time of garbage truck and send message to FMS, FMS can query historical records from database and send them to DAS. DAS can perform the procedure of data feature selection and data clustering methods to analyses travel time of garbage truck. The proposed methods can cluster the records of travel time and reduce variation for the improvement of travel time prediction. After predicting travel time and arrival time, the predicted information can be sent to user's MD.

The remainder of the paper is as follows. Section 2 proposes data feature selection and data clustering methods to analyse travel time of garbage truck for travel time prediction. The experimental results are presented and evaluated in Section 3. Finally, conclusions are given in Section 4.

2. Data Feature Selection and Data Clustering Methods

This study proposes data feature selection and data clustering methods which include seven steps to analyse and cluster the records of travel time. The details of procedure are illustrated as follows.

Step (1): Setting parameter

Step (1) sets the threshold of dependency ratio which is referred by cluster merging. When the dependency ratio between clusters is higher than threshold, the clusters should be merged.

Step (2): Selecting data feature

An unanalysed data feature can be selected for clustering data and calculating the dependency ratio between clusters. A data feature can have higher priority when the higher dependency ratio exists between clusters which are clustered in accordance with the selected data feature.

Step (3): Clustering in accordance with selected data feature

The records can be grouped in accordance with the selected data feature in Step (2), and the centroid of each group can be calculated. For instance, when the selected data feature is “weekday”, the records are divided into seven groups.

Step (4): Calculating the dependency ratio between clusters

Step (4) calculates the dependency ratio between each two clusters in accordance with the cumulative distribution function (CDF) of chi-square ($\chi 2$) distribution.

Step (5): Merging the clusters with a high dependency ratio and calculating the centroid of merged cluster

Step (5) compares each dependency ratio between each two clusters, and the two clusters with the highest dependency ratio are merged firstly. Then the centroid of merged cluster can be calculated.

Step (6): Checking unanalysed cluster

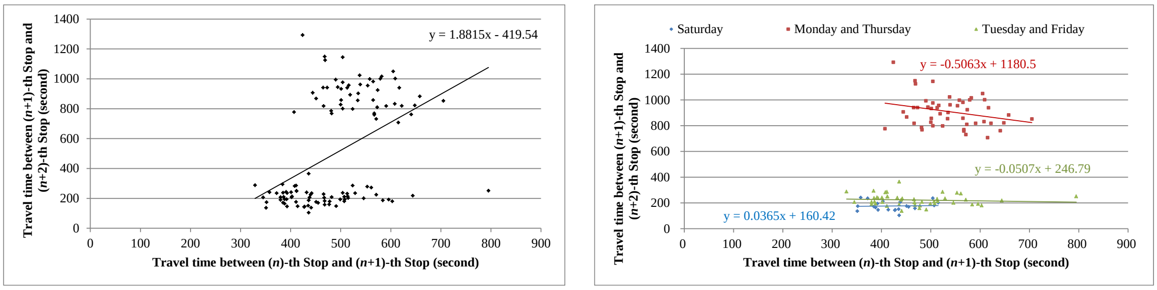
Steps (4) and (5) will be performed repeatedly until all clusters are analysed.

Step (7): Checking unanalysed data feature

Steps (2), (3), (4) and (5) will be performed repeatedly until all data features are analysed.

3. Implementation and Evaluation

In experimental environment, the records of travel time of garbage trucks in Hsinchu City during April and October 2014 were collected and analysed for the evaluation of proposed method. For travel time prediction, the multiple linear regression method was adopted to analyse the relation (i.e., slope and intercept) of the travel time between (n)-th Stop and (n+1)-th Stop and the travel time between (n+1)-th Stop and (n+2)-th Stop. Then this model could be used to predict the arrival time of (n+2)-th Stop when garbage truck arrived at (n+1)-th Stop [1]. The distribution of travel time was presented in Figure 1(a), and the accuracy of previous method was 16.73% as a result of the larger variation among the historical records of travel time. Therefore, this study used the proposed method to cluster the historical records of travel time into three groups (i.e., (1) Monday and Thursday, (2) Tuesday and Friday, and (3) Saturday) in accordance with weekdays. Then the previous method [1] was used to analyse the linear regression model of each cluster. The results which were showed in Figure 1(b) and Table 1 indicated the accuracy of proposed method was 85.97% for the improvement of previous method.



(**a**) (**b**)

Figure 1. The comparisons of travel time prediction

Table 1. The comparisons of travel time prediction

	The results of previous method [1]	The results of proposed method
Accuracy	16.73%	85.97%

4. Conclusions and Future Work

Due to the limitation of statistical methods and linear regression methods for travel time prediction, this study proposes a GTFMS and data feature selection and data clustering methods. The data feature selection and data clustering methods include seven steps: (1) Setting parameter, (2) Selecting data feature, (3) Clustering in accordance with selected data feature, (4) Calculating the dependency ratio between clusters, (5) Merging the clusters with a high dependency ratio and calculating the centroid of merged cluster, (6) Checking unanalysed cluster, and (7) Checking unanalysed data feature. The proposed methods can cluster the records of travel time and reduce variation for the improvement of travel time prediction. In experiments, the results showed that the accuracies of previous method and proposed method are 16.73% and 85.97%, respectively. Therefore, the proposed data feature selection and data clustering methods can be used to predict stop-to-stop travel time of garbage truck.

Author Contributions: The whole paper was done by Chi-Hua Chen.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. C.H. Chen, Y.T. Yang, C.S. Chang, C.M. Hsieh, T.S. Kuan, K.R. Lo, “The Design and Implementation of a Garbage Truck Fleet Management System,” South African Journal of Industrial Engineering, vol. 27, no. 1, pp. 32-46, 2016.
2. M.G., Karlaftis, E.I. Vlahogianni, “Statistical methods versus neural networks in transportation research: differences, similarities and some insights,” Transportation Research Part C: Emerging Technologies, vol. 19, no. 3, pp. 387-399, 2011.
3. C., Zhou, Z., Weng, C., Xu, Z. Su, “Integrated traffic information service system for public travel based on smart phones applications: a case in China,” International Journal of Intelligent Systems and Applications, vol. 5, no. 12, pp. 72-80, 2013.