Analyzing the regularities of passengers according to different time intervals via local electronic bus system data in Taiwan

Jing-Doo Wang[1] and Yao-Nan Lien[2] and Shin Hung Pan[3,*]

[1]Dept. of Computer Science and Information Engineering, Asia University, Taiwan.
Email: jdwang@asia.edu.tw
[1]Dept. of Medical Research, China Medical University Hospital, China Medical University.

[2] Dept. of Photonics and Communication Engineering, Asia University, Taiwan.
Email: yaonanlien@asia.edu.tw
[3] Dept. of M-Commerce and Multimedia Applications, Asia University, Taiwan.
Email: vincentpan@live.asia.edu.tw
[*] Corresponding Author

***Abstract*—It is interesting to extract and analyze the regularities of the passengers' behavior such that the government can provide public transportation support with quality control sufficiently and consistently . This study adopts the previous work that is a scalable approach based on Hadoop MapReduce programming model to extract maximal repeats from tagged sequential data and meanwhile to compute class frequency distribution of these maximal repeats, where the types of classes are derived from the tags attached with each of sequences according to users or domain experts in advance. In this study, experimental resources of sequential data contain the records of bus traffic data in 2015 and are authorized by the Taichung city government officially. Each record for one passenger includes two pair of timestamp and bus-stop representing for when and where that passenger on and off the bus, respectively, and the sequence of ordered bus-stops within his/her trip. Most of all, the types of tags for each of sequences include the types (CardType) and identification (CardID) of electronic card, the identification of bus route (RouteID) and the bus (BusID) that carried that passenger. Therefore, it is highly expected that the combination of these tags will provide many kinds of class frequency distribution and one can inspect the behaviors of buses and passengers from various points of view.**

Key words: Sequential data, Maximal Repeat, Bus Traffic, Hadoop, MapReduce

I. Introduction

For monitoring public transportation system in modern and metropolitan cities nowadays[5], it is usual that there are electronic devices deployed to collect the positions of vehicles and electronic tags attached individually to record the trips of passengers as time passes. Above information are usually collected as logs with timestamps like sequential data such that the government or transportation companies can monitor the flow of vehicles or passengers efficiently. With these huge amount of sequential data collected for a long time period, it is desired and attractive to have a scaleable approach to find out the repeats (or regularities) from these sequential data and then to inspect the schedules of vehicles and the behaviors of passengers such that the authorities can modify improper schedules and improve the efficiency of vehicle usages. Therefore, it is highly expected to have an efficient and scaleable approach to extract the repeats from these sequential data and to observe the frequency distribution of these repeats for further researches.

This study adopts the scalable approach [9] to extract maximal repeats [1] from a huge amount of tagged sequential data. In [9] for extracting maximal repeats, Wang proposed a scalable approach based on Hadoop MapReduce programming model [6] to overcome the computational bottleneck of using single computer with external memory[7], [8]. There were many experiments in diverse applications with a huge amount of tagged sequential data, such as textual data for trend analysis [9], genomic sequences for biomarkers identification [11], [13], timestamped gantry sequences for significant travel time intervals [12], the sequences of product traceability for quality control[10].

The remainder of this study is organized as follows. Section II briefly describes how to use the scalable maximal repeat extraction approach to compute the statistics of passengers according to different fixed-length time intervals. Section III gives the experimental results by selecting one bus trip for example. Section IV show the conclusions and discussions.

II. Methods

Figure 1 gives the conceptual diagram of processes to analyze the regularity within bus passenger trips. This paper adopts the scalable approach developed in [9], the rectangle with red dashed-box as shown in Fig. 1, to extract the behavior patterns of bus passengers within local electronic bus system in Taichung, Taiwan. The input of above approach consists of tagged sequential data that are the sequences of bus stops

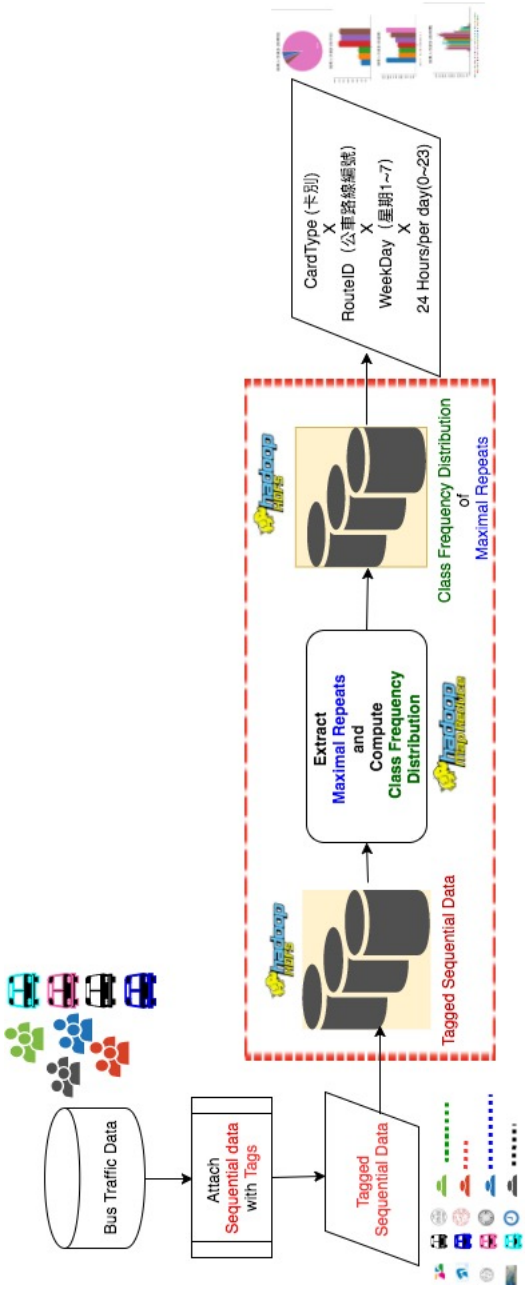

Fig. 1. The conceptual diagram of processes to analyze the regularity within bus passenger trips

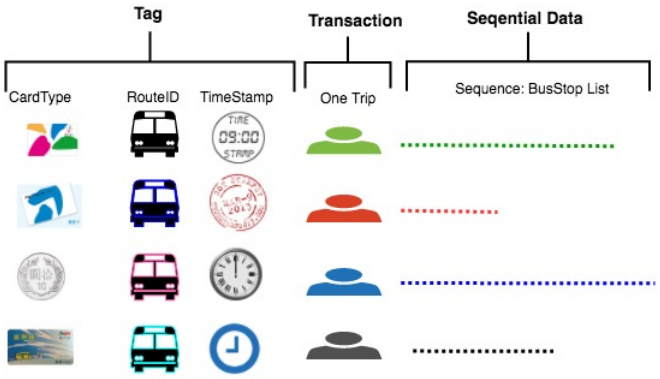

Fig. 2.

資料編號，路線，日期，星期，公車車號，票種，卡號，上車時間，下車時間，上車站#.....#下車站

5cb58a48-280f-4801-b376-5162dccff31e,159,2015-01-09,5,419-FZ,中市敬老,0000167A,"2015-01-09 11:33:30.000","2015-01-09 12:06:28.000",英才西屯路口#科學博物館(臺灣大道)#公益公園(忠明南路)#忠明南路向上路口

22234abd-c681-4e2b-a170-89094fbc97e5,58,2015-01-30,5,546-FQ,全票,0000515C,"2015-01-30 15:47:00.000","2015-01-30 15:51:00.000",臺中公園(雙十路)#干城站#臺中火車站

Fig. 3. Example of tagged sequential data: one fields list with two sequences with tags in the front.

within each of passenger trips in ”Bus Traffic Data”; each of those sequences are attached with the tags including, ”CardType”, ”RouteID” and ”TimeStamp”, as shown in Fig.2; the output of that approach is the class frequency distributions of maximal repeats extracted from those tagged sequences, where the classes are derived from the tags. Note that above scalable approach developed in [9] was based on Hadoop MapReduce programming model

To illustrate the input of sequences, for example as shown in Fig.3, the first row contains the field names of one record, and the last two rows are two records with values in order as corresponding field names in the first row. Note that the last item of each sequence is the list of bus stops, separated by ”#”, of one passenger trip. The output generated by the maximal repeat extraction , for example as shown in Fig.4, in which contains ”One Maximal Repeat” with meta data and class frequency distributions that separated by ”##”. The details of these class frequency distribution are illustrated in Section III.

III. Experimental Results

In this study, the experimental resources are collected from the Taichung Bus Electronic Tag (e-Tag) system during the

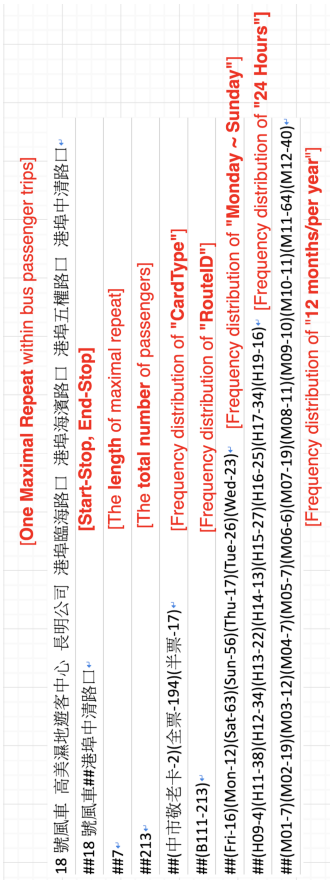

Fig. 4.

Fig. 5. A web site (http://titda.asia.edu.tw/) constructed for illustration.

year ”2015”. Recently, there were studies based on these data [3], [2], [4]. To solid the contribution of this work, there is a web site (http://titda.asia.edu.tw/) constructed for illustration. For example, as shown in Fig.5, one can have a query with the input as ”RouteID=300” and two bus stops, the beginning (Providence University) and the ending stops (Taichung Train Station), then the web system will output the statistics of the number of passengers in that trip according to different types of classes as follows.

Fig.6, first of all, gives the number of passengers according to the ”CardType” in Fig.5. It is obvious that the type

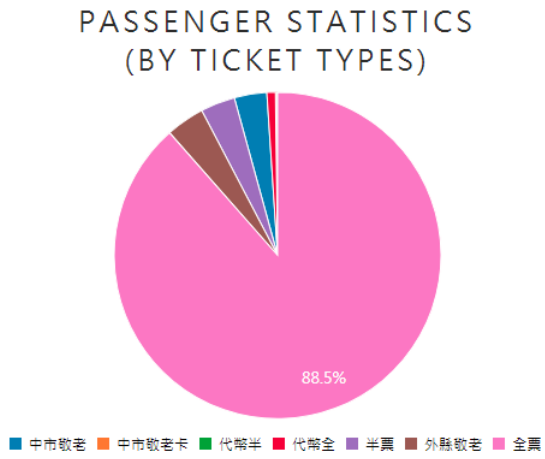

Fig. 6. The number of passengers according to the ”CardType” in Fig.5

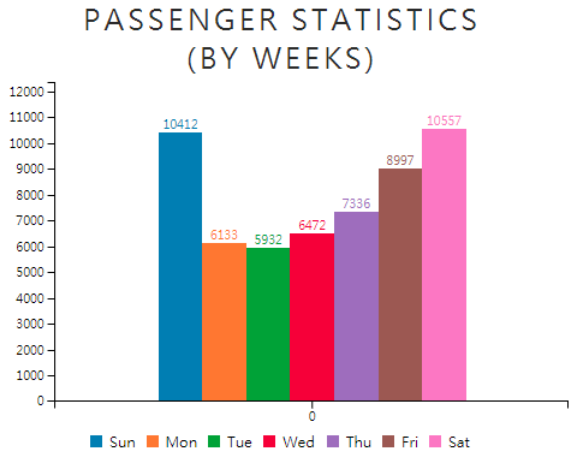

Fig. 7. The number of passengers according to the ”WeekDay” in Fig.5

”CardType”=”全票”(full fare) achieved the highest percentage ”88.5%”. Secondly, Fig.7 shows the number of passengers according to the ”Weekday”. One can find that the numbers of passengers in the weekend (Friday, Saturday and Sunday) were the top 3 within the weekday comparison. Similarity, Fig.8 and Fig.9 give the numbers of passengers according to the types of classes as ”24 hours/per day” and ”Month”, respectively. One can observe the frequency distribution of passengers as to different fixed-length time intervals. For domain experts, it is desired to combine different tags as new class to inspect specific frequency distribution on purpose in the future if necessary.

IV. Conclusion and Discussion

This study provides fundamental experiments to survey the statistics of passengers in the local electronic bus system data

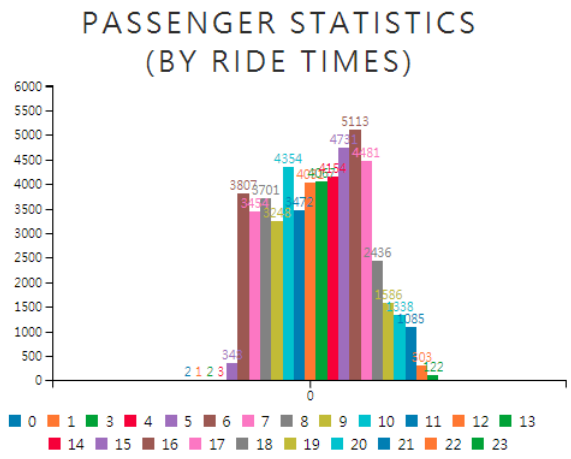

Fig. 8. The number of passengers according to the "24 hours/per day" in Fig.5

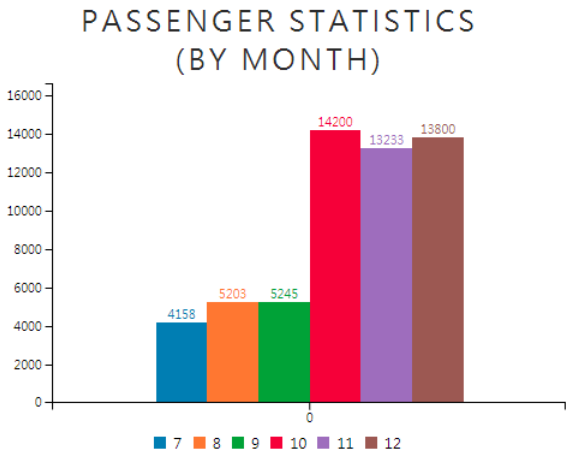

Fig. 9. The number of passengers according to the "Month" in Fig.5

in Taichung, Taiwan. With the scalable approach [9] to extract maximal repeats and meanwhile to computer class frequency distributions of these repeats from tagged sequences, consisting of bus stops list of each of passenger trips, where the types of classes are derived from the tags, e.g. "CardType", "RouteID" and "Timestamp", one can observe the frequency of passengers according to what types of classes are selected for observation.

Indeed, there are rooms to further extend the strength of this study by combining different tags arbitrarily, e.g. "Cardtype+Weekday" or "RouteID+24 hours/per day", to define specific class frequency distribution of patterns (bus stops list) for observation on purpose if necessary. This study can provide diverse frequency distribution of bus passengers for observation as long as someone desires. These experimental results are not allowed to be opened globally and publicly due to these information of electronic e-tags gathered from bus passenger trips is authorized by Taichung local government so far. It is expected that the strength of this study can have more influences for improving the efficiency of bus traffic transportation or for helping the government or bus company to adjust their bus schedule if above restriction of using these e-tags data can be removed and more public opinions or experts are involved with.

V. Acknowledgement

The database used in this article is called Taichung City Smart Transportation Big Data Database which is provided by the Bureau of Transportation, Taichung City Government. We would like to give special thanks to Taichung City Government and Asia University for their immense supports.

References

[1] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences : computer science and computational biology*. Cambridge University Press, 1997.

[2] Cheng-Yuan Ho, Chih-Chun Huang, Yi-Jhong Huang, Wei-Tse Lin, and Shao-Lun Wu2. The Ridership Analysis of Taichung Bus Route - A Case Study of Wufeng District. In *In Proceedings of The 2017 International Forum on Taichungology*, pages 129–140, 2017.

[3] Cheng-Yuan Ho, Shu-Chuan Liao, Yao-Nan Lien, and Yi-Chuan Wang. Health-oriented Welfare Policy for the Elderly: A Case Study of Senior Citizens' Bus-Riding Frequency in Taichung City. In *In Proceedings of The 2017 International Forum on Taichungology*, pages 191–202, 2017.

[4] Cheng-Yuan Ho and Yao-Nan Lien. A Study on Applying the Big Data of Electronic Ticket to Construct the Passengers' Patterns of Taichung City Bus. In *In Proceedings of The 13th Taiwan Conference on Software Engineering*, pages 98–98, 2017.

[5] E. Macioszek and G. Sierpiński. *Recent Advances in Traffic Engineering for Transport Networks and Systems: 14th Scientific and Technical Conference "Transport Systems. Theory & Practice 2017" Selected Papers*. Lecture Notes in Networks and Systems. Springer International Publishing, 2017.

[6] Yu Shyang Tan, Jiaqi Tan, Eng Siong Chng, Bu-Sung Lee, Jiaming Li, Susumu Date, Hui Ping Chak, Xiong Xiao, and Atsushi Narishige. Hadoop framework: impact of data organization on performance. *Software: Practice and Experience*, 43(11):1241–1260, 2013.

[7] Jing-Doo Wang. An external memory approach to compute the statistics of maximal repeats across classes from whole genome sequences. In *2005 National Computer Symposium, Taiwan, R.O.C*, pages BIC1–2, 2005.

[8] Jing-Doo Wang. External memory approach to compute the maximal repeats across classes from dna sequences. *Asian Journal of Health and Information Sciences*, 1(2):276–295, 2006.

[9] Jing-Doo Wang. Extracting significant pattern histories from timestamped texts using mapreduce. *The Journal of Supercomputing*, pages 1–25, 2016.

[10] Jing-Doo Wang. A novel approach to improve quality control by comparing the tagged sequences of product traceability. In *The 3rd International Conference on Inventions*, 2017.

[11] Jing-Doo Wang, Wen-Ling Chan, Charles C.N. Wang, Jan-Gowth Chang, and Jeffrey J.P. Tsai. Mining distinctive dna patterns from the upstream of human coding and non-coding genes via class frequency distribution. In *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2016)*, 2016.

[12] Jing-Doo Wang and Ming-Chorng Hwang. A novel approach to extract significant patterns of travel time intervals of vehicles from freeway gantry timestamp sequences. *Applied Sciences*, 7(9), 2017.

[13] Jing-Doo Wang, Yi-Chun Wang, Rouh-Mei Hu, and Jeffrey Tsai. Extracting the co-occurrences of dna maximal repeats in both human and viruses. In *The 17th annual IEEE International Conference on Bioinformatics and Bioengineering (BIBE2017)*, 2017.