IET Smart Cities

Research Article

Online Web query system for various frequency distributions of bus passengers in Taichung city of Taiwan

eISSN 2631-7680
Received on 14th March 2020
Revised 8th July 2020
Accepted on 17th July 2020
doi: 10.1049/iet-smc.2020.0017
www.ietdl.org

Jing-Doo Wang[1,2], Shin-Hung Pan[3] ✉, Cheng-Yuan Ho[1], Yao-Nan Lien[1], Shu-chuan Liao[4], Achmad Nurmandi[5]

[1]*Department of Computer Science and Information Engineering, Asia University, Taiwan*
[2]*Department of Medical Research, China Medical University Hospital, China Medical University, Taiwan*
[3]*Department of M-Commerce and Multimedia Applications, Asia University, Taiwan*
[4]*Department of Social Work, Asia University, Taiwan*
[5]*Department of Government Affairs and Administration, Universitas Muhammadiyah Yogyakarta, Indonesia*
✉ *E-mail: vincentpan@live.asia.edu.tw*

Abstract: It is highly desirable that traffic controllers or city residents can discern regular patterns and promptly detect irregularities or abnormal events in a public transportation system. This study proposes a web-based information system that allows users to study the travel behaviour of bus passengers from various perspectives. The system uses data from the comprehensive set of Taichung City Bus Riding Records between 2015 and 2016. However, it can provide the same functionality to any other similar bus transportation system by using the appropriate data. It should be emphasised that the system can provide the frequency distributions not only of passenger trips between two stops but also of the passenger volume for a given segment of any route. Owing to the increased computational and storage-capacity requirements of the proposed system, the scalable Hadoop MapReduce programming model was used. Furthermore, bus companies can use the system to design better service plans, such as more flexible bus schedules and more convenient routes, to meet passenger demand as well as reduce operation cost and energy consumption. The authors believe that the proposed system can make a valuable contribution to public welfare.

1 Introduction

The efficiency of a transportation system is usually analysed by selecting specific or representative routes and then computing statistics of passenger trips on these routes for further investigation or comparison. However, in a modern city, to analyse traffic bottlenecks or to monitor the public transportation system globally, it is essential to inspect or compare the frequency distributions of passenger trips on all routes from various perspectives as systematically as possible. By observing the distribution of passenger trips in various fixed-length time intervals, for example, 24 h per day, day of the week, month, or year, one may discern regularities or detect abnormalities, thereby aiding in improving the quality of transportation services. It is not easy to develop a public query system whereby one can instantly inspect the statistics of passenger trips between two arbitrary stops on any route.

1.1 Bus transportation system in Taichung city

Taichung city, located in the middle of Taiwan, became a municipality with over two million residents after merging with Taichung County in 2010. The website ‘The public transportation information of Taichung’ [1] has provided route maps for the public transportation networks since 2015. Taichung city had >100 bus routes when this study was launched. For example, Fig. 1 provides integrated information regarding 11 bus routes (300âĂŞ310), which were individually serviced by four bus companies. It should be noted that routes 309 and 310 were not available when this study was launched. Fig. 2 shows the bus stop ‘Providence University’, which is used by numerous students. Fig. 3 shows a bus for route 300 with license plate number (BusID) KKA-1063 on Taiwan avenue in Taichung. To collect electronic records of all passenger trips automatically, as shown in Fig. 4, all passengers should swipe their card when they embark and disembark. It should be noted that these routes may have overlapping parts, as shown in the middle of Fig. 1, where all 11 routes have a common segment on Taiwan avenue. It would be interesting to inspect the variations in the frequency of passenger trips for arbitrary segments on these routes.

1.2 Frequency distributions of passenger trips and passenger volume

The concept of this study is illustrated in Fig. 5, which shows six-passenger trips (P-1–P-6) involving stops S-1–S-8, through which the routes 300, 301, 303 and 304 passes. In trip P-2, for example, a passenger boards bus 04, belonging to route 304, at stop S-2 and disembarks at stop S-6, using an EasyCard [3] or another card (see the Appendix) to pay the fee for that trip. Similarly, in P-4, P-5 and P-6, passengers board bus 01, belonging to route 300, at stop S-4 and disembark at stop S-8. Using the ordered stops as the indices of a two-dimensional (2D) matrix (Table 1), it is easy to transform P-1–P-6 into four cells in Table 1(a), whereby the frequency of the trips can be calculated. Generally, as the record of a trip contains two stops to indicate where a passenger embarks and disembarks on a certain route, it is straightforward to obtain the frequency (number) of passenger trips (as in the example in Table 1(a)) by scanning all the records and accumulating the number of trips using a 2D matrix. This computation can be easily carried out using a structure query language (SQL) command in a traditional relational database system.

To compute passenger volume, that is, the total number of passengers passing through any two given stops for all routes, additional effort is required, particularly if the lists of ordered stops for the routes are not the same, and if some of these lists partially overlap. To obtain passenger volume, one should trace all the stops through which each passenger trip passes. For example, as shown in Table 1(b), trip P-2 results in a series of value 1 between two consecutive stops from S-2 to S-6. Similarly, P-4, P-5 and P-6 generate a series of value 3 between five consecutive stops from S-4 to S-8. The frequency of passenger volume can be obtained by summing all the cells in the same column, as shown at the bottom

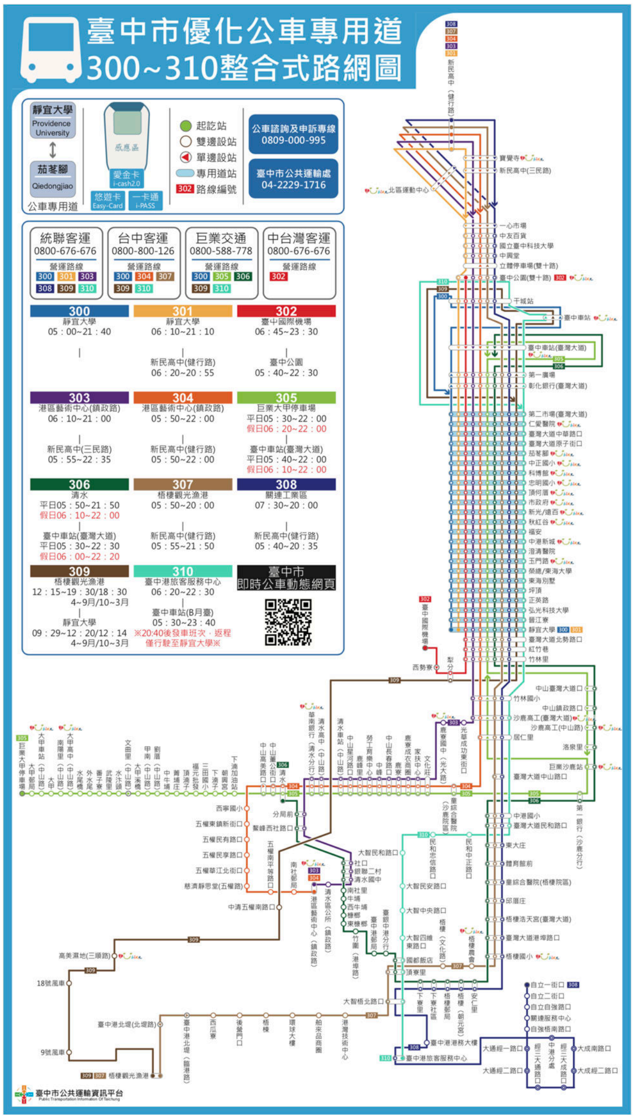

Fig. 1 *Map of bus route 300–310 in Taichung city [2]*

Fig. 2 *Bus stop 'Providence University' at Taiwan avenue in Taichung city*

of Table 1(b). Unlike in the case of passenger-trip frequency, as shown in Table 1(a), it is desirable to detect or identify the hottest interval of consecutive stops (S-4–S-6), through which all six passenger trips pass. Generally, such an interval is not easily detected using a traditional 2D matrix (such as Table 1(a)). Moreover, bus routes are usually designed as different ordered bus stops to cover a service area as large as possible, and occasionally, as shown in Fig. 1, there are intersections or overlapping segments on these routes when transportation efficiency is considered. That is, it is considerably difficult to use a totally integrated 2D table, such as Table 1, for the frequency of passenger volume because the bus stops lists are not the same for all routes.

1.3 Contribution of this study

It is highly desirable to construct a web query system to provide users with online services so that domain experts or city residents can observe various class-frequency distributions of passenger trips between any two bus stops on an arbitrary route. This website can provide various frequency distributions of passenger trips or passenger volume for arbitrary intervals on any route. Indeed, this study is an extension of [4, 5]; however, it has more robust experimental resources involving two-year records from 2015 to 2016 and provides various interactive 3D data visualisation interfaces through the ECharts visualisation library [6].

With this web query system, in addition to domain experts or city government, city residents can inspect the frequency distributions corresponding to segments on routes with which they are familiar, and then they can use this information in their trip planning, and provide suggestions to the city government or the bus company. This may be critical to the city government or the bus company, but it may reflect some hidden problems from the residents. This web query system can assist the residents, the city government or the bus company in globally and systematically inspecting the bus services for all routes. Furthermore, the processes for constructing the web query system (in terms of both hardware and software) are described in detail, so that this system can be reproduced in other modern cities with complicated public transportation systems, provided that electronic records of all passenger trips and the stops for all routes are available.

The remainder of this paper is organised as follows. In Section 2, the system diagram, which consists of three subsystems, is provided, and then each of these subsystems is described in detail. In Section 3, the experiments and results are presented. In Section 4, a discussion is provided, and future directions are indicated. Finally, Section 5 concludes the paper.

2 Method

The aim of this study is to construct a web query system for various frequency distributions of bus-passenger volume for any two stops on a route according to fixed-length time intervals, which are used as bin-counting units for the frequency distributions and are defined in advance, for example, year, month, days of the week, and 24 h per day. The computation of the frequency distributions is based on the classes (types) of information (tags) associated with bus passenger trips, including bus routes (Route), types (Ticket Type) of electronic passenger cards, and two-timestamps for embarking and disembarking. For comparison, these frequency distributions include not only the number of passenger trips corresponding to embarkation or disembarkation on a certain route but also the number (volume) of those passing through two stops on all routes. In addition to the frequency distributions based on different time intervals, one also can observe various class frequency distributions of passenger volume, where the class types are derived from the tags representing trip features, such as the type of electronic cards and the route identifier.

Fig. 6 shows the conceptual diagram of this study. It consists of three subsystems, each surrounded by a dashed rectangle. The blue-dashed rectangle (upper right) corresponds to the subsystem that generates tagged, sequential bus stop lists for all bus passenger volumes using the ordered stop names for each bus route. The details of this subsystem are given in Section 1. The black-dashed rectangle (upper left) corresponds to the subsystem that uses the aforementioned stop lists as an input and then computes various class frequencies of the passenger volume between two arbitrary stops on a route. Owing to the increased computational and storage-capacity requirements for performing these processes, a scalable approach was adopted [7, 8]. Specifically, we used the Hadoop MapReduce programming model [9] with a cluster of computing nodes. In Section 2.2, we describe the process of extracting maximal repeats from these tagged, sequential bus stops and computing various class frequencies of specific segments on a bus route. After these frequency distributions are imported into a relational database (MS SQL server), in Section 2.3, we demonstrate how web-interface queries for specific class frequency distributions of segments on bus routes can be made by using ECharts [6], thus providing a visual web service.

Fig. 3 *One bus with 'BusID' as 'KKA-1063' for Route '300'*

Fig. 4 *Passenger should swipe his/her own card twice per trip when he/she gets on/off the bus*

2.1 Generating ordered bus stop lists of passenger volume attached with tags

The subsystem marked with the blue-dashed rectangle in Fig. 6 is used to transform each passenger trip into a tagged (or labelled) ordered bus-stop list. These lists are used as the input for the next step described in Section 2. The distinct tag values or their combinations can be used as the units (classes) for bin-counting to compute the frequency distributions of segments on bus routes. As shown in Fig. 7, the classes of these tags may be different for electronic cards (Ticket Type), bus routes (Route), or fixed-length time intervals (year, month, day of the week, 24 h per day) derived from the timestamps. In addition to the names of two stops stored during embarkation and disembarkation for each passenger trip, ordered bus-stop lists are generated as consecutive names of all stops from embarkation to disembarkation.

As several bus companies operate in Taichung, as shown in Fig. 1, a stop may have different names. However, for consistent extraction from the tagged bus-stop lists, it is important for each stop to have a unique name or identifier. Accordingly, the longest common subsequence approach [10] was adopted to change each ambiguous stop name with the most similar one whose name was unified officially.

Fig. 5 illustrates the principle of this study: there are six passenger trips (P-1–P-6) on four bus routes (Route) (300, 301, 303 and 304). For example, in P-2, a passenger boards bus 04, belonging to route 304, and has his/her electronic card scanned upon embarkation at stop S-2 and disembarkation at stop S-6, respectively. These passenger trips are transformed into six ordered bus-stop lists tagged with 'Ticket Type', 'Route' and 'TimeStamp' for embarkation/disembarkation (Fig. 7).

2.2 Computing class frequency distributions

2.2.1 Maximal repeat: This study adopts maximal repeats as the units for computing class frequency distributions of passenger volume. To capture regularity in sequential data, that is, the ordered stop lists of all passenger trips, it is essential for further analysis to extract repeats that are consecutive sequences and appear at least twice. The definition of maximal repeats was proposed in [11] to address problems in genomic sequences. Intuitively, a maximal repeat cannot always be a substring of another repeat in all sequential data. However, a repeat may be omitted in the computation of frequency distributions if it is always a substring of a maximal repeat because the statistics of a repeat are always the same as those of the maximal repeat. For example, the segments S-4, S-5 and S-6 in Fig. 7 are a maximal repeat. However, the segments S-4, S-5 and S-5, S-6 are not maximal repeats because they are substrings of S-4, S-5 and S-6 in the ordered bus-stop lists in Fig. 7. That is, these two segments have the same statistics as

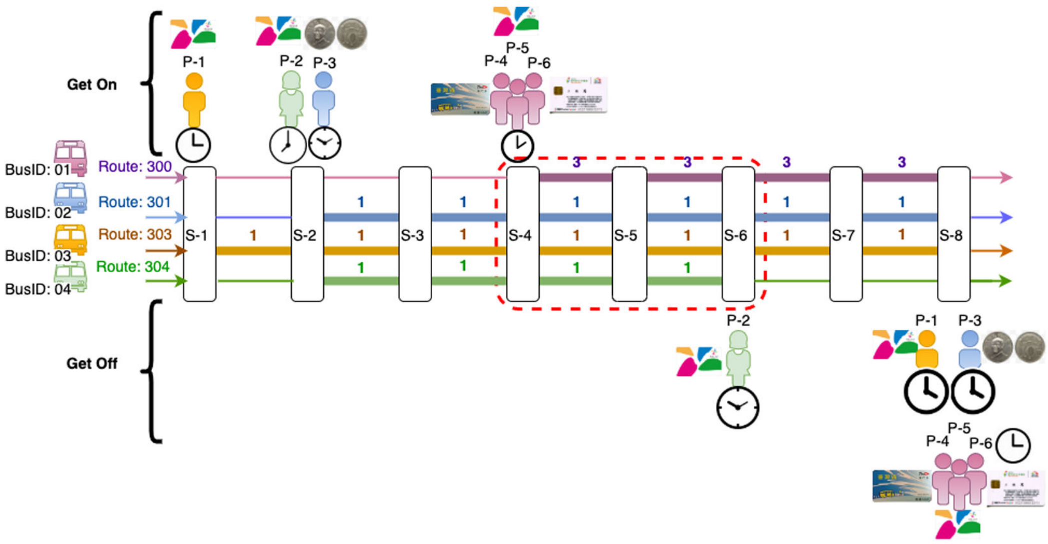

Fig. 5 *Example of six-passenger trips within four bus routes where the hottest interval is marked with a red-dashed rectangle*

Table 1 (a) Passenger trips (b) passenger volume derived from Fig. 5

		End stop							
		S-1	S-2	S-3	S-4	S-5	S-6	S-7	S-8
(a) frequency of passenger trips for getting on/off stop									
start stop	S-1	X							1
	S-2		X				1		1
	S-3			X					
	S-4				X				3
	S-5					X			
	S-6						X		
	S-7							X	
	S-8								X
(b) passenger volume for passing through two given stops									
		end stop							
		S-1	S-2	S-3	S-4	S-5	S-6	S-7	S-8
start stop	S-1	1	1	1	1	1	1	1	1
	S-2		2	2	2	2	2	2	2
	S-3								
	S-4				3	3	3	3	3
	S-5								
	S-6								
	S-7								
	S-8								
		1	3	3	6	6	6	5	5

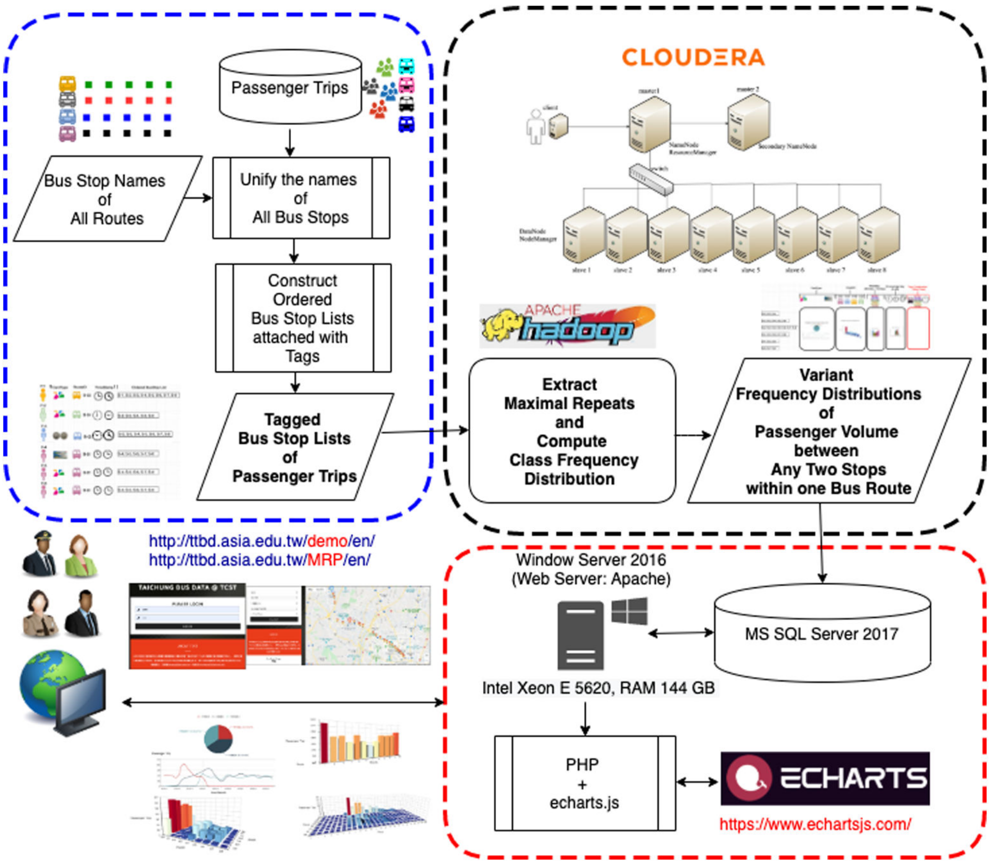

Fig. 6 *Conceptual diagram of constructing web query system for various frequency distributions of bus passenger volume*

those of S-4, S-5 and S-6, and therefore they can be omitted to reduce computation cost.

2.2.2 Extracting maximal repeats and computing their class frequency distributions: As shown in the upper right of Fig. 6, the subsystem with the black-dashed rectangle adopts a scalable maximal repeat extraction approach [7, 8] that is based on distributed computing using the Hadoop MapReduce programming model [9] and is capable of handling a large amount of sequential data. To make this paper self-contained, the processes of extracting maximal repeats from the tagged ordered stop lists of all passenger trips are briefly described; however, one can find a detailed description in [7].

Conceptually, the aforementioned scalable extraction approach consists of three steps. In the first step, each tagged and ordered stop-list is transformed into all of its suffixes with their corresponding tags, where one ordered list with n items (stops) generates n suffixes [11]. Subsequently, these suffixes are sorted

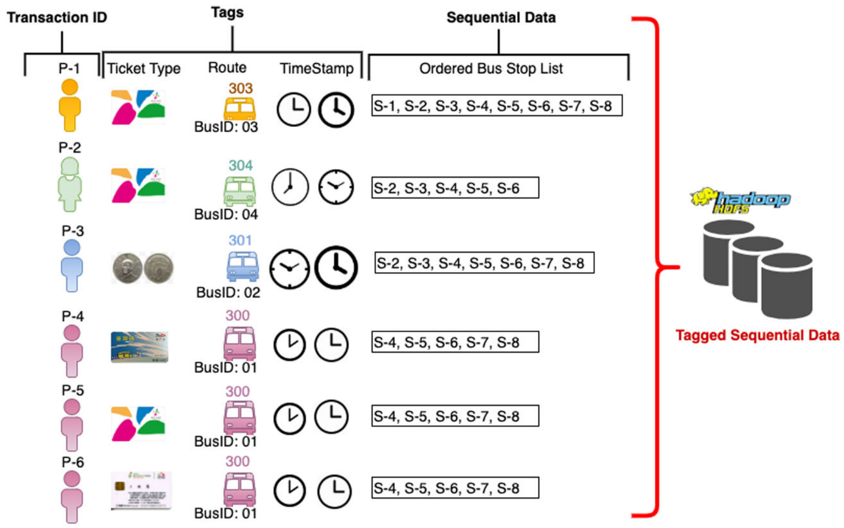

Fig. 7 *Six ordered stop lists with tags constructed according to the six-passenger trips in Fig. 5*

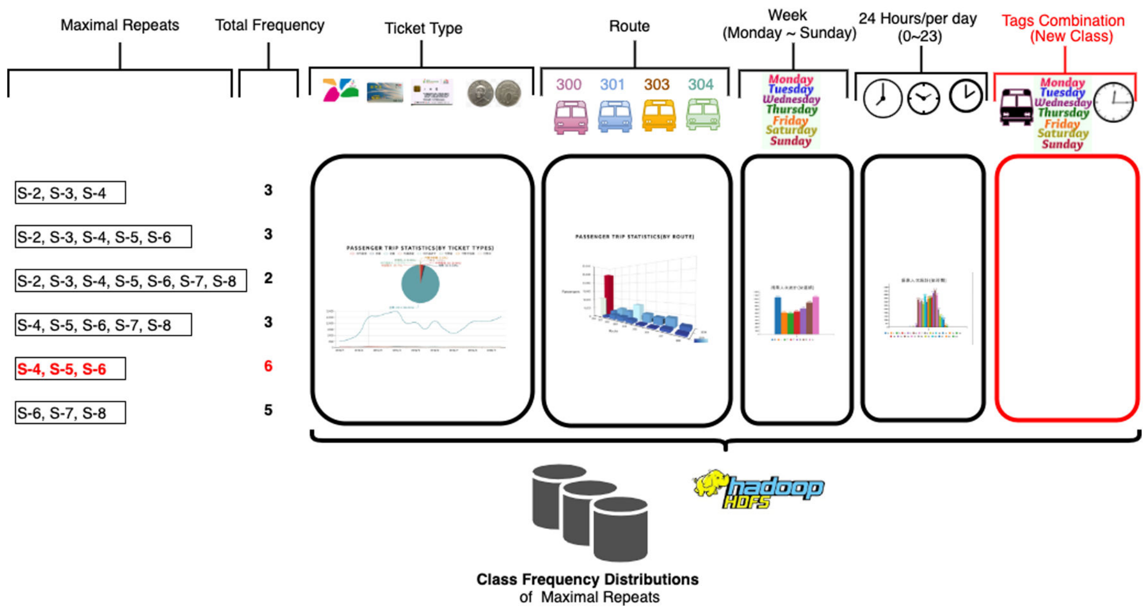

Fig. 8 *Six maximal repeats extracted from tagged stop lists obtained in Fig. 7 and their class frequency distributions*

and scanned so that the longest common prefix of two adjacent sorted suffixes may be extracted as a candidate maximal repeat; the statistics of the class frequency distribution of this candidate repeat are accumulated using a stack with push/pop operations, where the class types are the tags of the original list (passenger trips). The repeats extracted in the first step are termed as candidate maximal repeats with 'right' verification. In the second step, the processes are similar to those in the first step, but the items of all lists are ordered in reverse. The repeats obtained in the second step are candidate maximal repeats with 'left' verification. In the final step, a repeat is considered maximal if it has 'right' verification and its reverse has 'left' verification. This approach is scalable for the following reasons. One is that the computationally intensive task of suffix sorting is performed automatically by the Hadoop system [12]. Another is that the scanning of sorted suffixes can be smoothly partitioned into smaller tasks by using the prefix items of the sorted suffixes as keys in the MapReduce programming model.

This scalable approach can extract maximal repeats from tagged, sequential bus-stop lists (generated in Section 1) and compute the class frequency distributions of these repeats, where the classes are derived from the tags. As the class types can be arbitrary tag combinations (as is expected) the dimension of these distributions can be quite high to capture the passenger volume for two arbitrary stops on a bus route. Owing to the scalability requirement regarding both computational resources (for extracting maximal repeats from a large number of tagged bus-stop lists) and storage capacity (for storing these repeats and their frequency distributions), this subsystem was implemented on a Hadoop cluster.

2.2.3 Various class frequency distributions: Fig. 8 shows six maximal repeats extracted from the ordered bus lists in Fig. 7, and various frequency distributions of these repeats according to the classes derived from previous tags. It should be noted that the classes are user-defined and may be any combination of tags. Indeed, these repeats can be estimated as segments between two stops on a bus route. Then, given such a segment, various frequency distributions can be obtained to observe or compare passenger-volume variations.

2.3 Establishing web query system with visual interfaces

As shown in the red-dashed rectangle in Fig. 6, the webserver was a computer running Windows Server 2016 with database server MS SQL Server 2017; the computer had two Intel Xeon E 5620 CPUs at 2.40 GHz and 144 GB random access memory (RAM). Moreover, this study provided a data visualisation interface for observing various class frequency distributions, integrating PHP with echarts.js (javascrpt) to request on-line services from ECharts [6].

For application in Taichung city, an interactive web query system, located at http://ttbd.asia.edu.tw/MRP/en/, as shown in

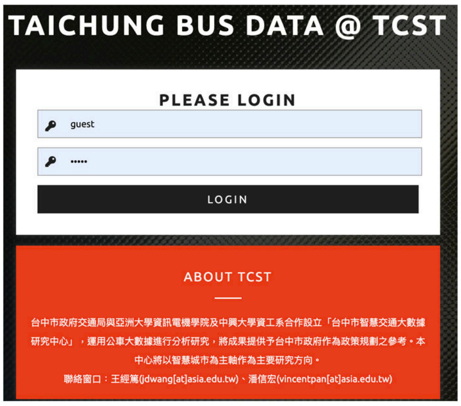

Fig. 9 *Web query system for bus-passenger frequency distribution (2015–2016) in Taichung, Taiwan (http://ttbd.asia.edu.tw/MRP/en/)*

Table 2 Statistics of passenger trips in Taichung city of Taiwan

	Number of passenger trips	Number of bus routes	Number of distinct bus stops	
			Get-on	Get-Off
2015	82,819,925	280	4526	4536
2016	118,775,493	183	2966	2994
—	201,595,418	—	—	—

Fig. 10 *Example of tagged and ordered bus-stop list*

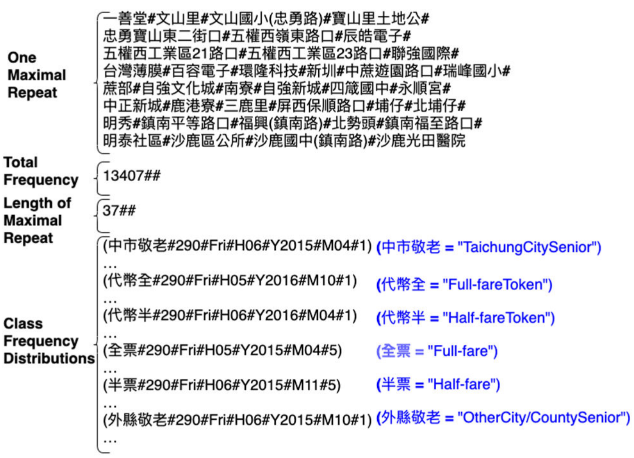

Fig. 11 *Example of maximal repeat and its class frequency distributions*

Fig. 9, was constructed, and a guest account was set up for public testing.

For comparison with traditional statistics of passenger trips based on two specified stops (embarkation/disembarkation), a similar web interface is also provided at http://ttbd.asia.edu.tw/demo/en/. Intuitively, the latter is based on the computation in Table 1(a), whereas the former on that in Table 1(b).

3 Experimental results

3.1 Experimental resources

As described in Table 2, in the experiments, we used 201,595,418 passenger trips from 2015 to 2016, authorised by the Taichung city government. It is easy to observe that the number of bus routes and distinct bus stops in 2015 was larger than that in 2016 because, in 2015, the bus routes were adjusted to improve the efficiency of the transportation system.

After the names of all bus stops were unified, each bus-passenger trip was transformed into a tagged, ordered bus-stop list. Fig. 10 shows an example of such a list (from embarkation to disembarkation), containing nine consecutive stops separated by the symbol '#'. The list is pre-tagged with 'Passenger Trip ID' = 5cb58a48-280*, Route = 159, Date = 2015-01-09, Week = 5, BusID = 419-FZ, 'Ticket Type' = 'Elder in Taichung', CardID = 000016* and two timestamps for embarkation/disembarkation. The tag 'Passenger Trip ID' is individually assigned to each passenger trip; 'Route' refers to a specific road map with sequential bus stops (for drivers as well as passengers); 'Date' and 'Timestamp' refer to the date and time for a certain trip, respectively; 'Week' refers to the days of the week. For example, 'Week = 5' implies Friday; 'BusID' is used to identify buses. That is, there may be several buses (BusID) running on the same route (Route) simultaneously during rush hours. The type of electronic card 'Ticket Type' is used to determine the fare paid by a passenger per trip. Each electronic card 'CardID' is valid if its account is loaded with money. Tags in red in Fig. 10 are used to compute class frequency distributions of maximal repeats, as described in Section 2.

3.2 Class frequency distributions of maximal repeats

In this study, as described in Table 2, 1364,272,545 maximal repeats were extracted from the passenger trips, and various class frequencies of these repeats were extracted by the scalable approach described in Section 2. A maximal repeat can be considered as an ordered stop list between two arbitrary stops on a bus route. The size of the database on the MS SQL server 2017 for storing these repeats and their class frequency distributions was ∼200 GB, with an additional 104 GB for indexes. Fig. 11 shows an example of a maximal repeat extracted from tagged bus-stop lists that contained 37 stops and appeared 13,407 times totally. For example (Full-fare"#290#Fri#H05#Y2015#M04#5) in Fig. 11 indicates that a maximal repeat appeared five times in the case of passenger trips using 'Ticket Type' = Full-fare on Route = 290 between 5 am and 6 am (H05) when 'Week' was Friday (Fri) in April 2015 (Y2015#M04) (Fig. 12).

3.3 Web querying system for various class frequency distributions

For application in Taichung city, as shown in Fig. 9, one can log into the web query system at http://ttbd.asia.edu.tw/MRP/en/ using the guest account to observe various frequency distributions of passenger trips between any two stops on a bus route. With input parameters as shown in Fig. 13, one can obtain the class frequency distributions for two given stops on route 300 during the year 2016. The related observations are described in Sections 3.3.1–3.3.3 (Fig. 14).

3.3.1 Ticket type versus passenger volume: To promote the use of public bus transportation with electronic card (see the Appendix), the Taichung city government proposed policy in 2012, called '8 km free', whereby any passenger trip with the electronic card was free if its distance was <8 km. Accordingly, it is quite

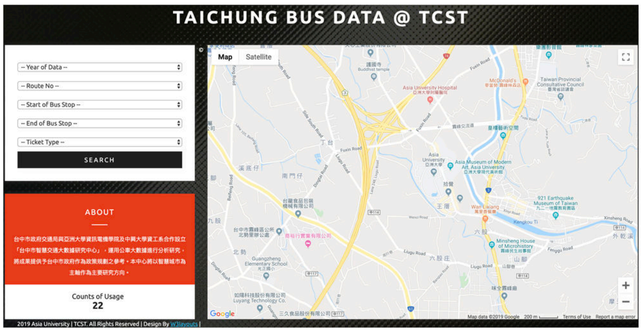

Fig. 12 *Input parameters for query system with Google Maps around Asia University in Taichung city of Taiwan*

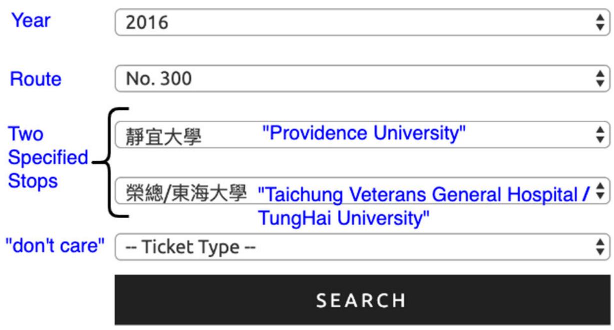

Fig. 13 *Example of parameters with two specified stops for Route = 300 in 2016*

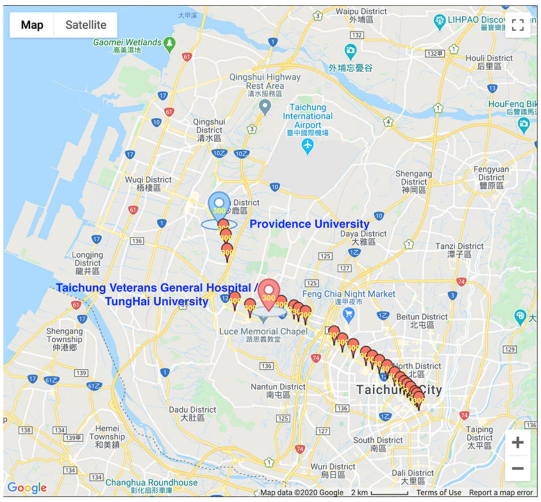

Fig. 14 *Google Map of route 300 with two specified stops in Fig. 13*

interesting to observe the variations of class frequency distributions from any segments on all routes. Fig. 15 shows the monthly frequency distributions of passenger trips according to Ticket Type under the conditions described in Fig. 13. The lowest passenger volume (10,031) when Ticket Type was Full-fare appeared in ‘2016/7’.

3.3.2 ‘Month’, ‘Week’ and ‘24 Hours’ versus passenger volume: As shown in Figs. 16–18, the frequency distributions of the passenger volume for route 300 can be obtained according to different fixed-time intervals, namely, ‘Month’, ‘Week’ and ‘24 Hours’. In Fig. 16, it is seen that there were two troughs around the months ‘2016/4’ and ‘2016/7’ related to the spring (April) and summer vacations for students in Taiwan. Fig. 18 shows that, among the days of the week, the peak of passenger volume occurred on Friday. Fig. 18 indicates the rush hours during a day were between 4 pm and 6 pm, as students or workers usually return to their residence during that time. These observations could provide clues or hints to the local bus company to adjust the time schedule for route 300 and to provide bus services more efficiently.

3.3.3 ‘Elder’ versus passenger volume: During the time period from 2015 to 2016, residents of Taichung city over 65 years of age could apply for a specific type of electronic card, namely, ‘Elder in Taichung’. Subsequently, the Taichung city government would automatically upload 1000 NT every month. Children between the ages of 6 and 12 have a 50% discount with ‘Ticket Type’ = ‘half-price’. People over 65 years of age may apply for the ticket discount when they use public transportation in Taiwan. To observe various frequency distributions for elderly passengers, as shown in Fig. 19, one can retain the same parameters as in Fig. 13 but change ‘Ticket Type’ to ‘All of the Elders’ and set the parameters for ‘Year’ and ‘Route’ to ‘don't care’ in the query. That is, one can capture regular patterns regarding the travel behaviour of elderly people, thus facilitating social welfare workers and the Taichung city government to provide more suitable public transportation services for this age group (Fig. 20).

Fig. 21 shows the percentage for the three types of ‘Ticket Type’ from ‘2015/7’ to ‘2016/12’; one of these types was zero in ‘2016/2’. It is interesting to determine the cause of this. Moreover, Figs. 22 and 23 show the statistics of the routes passing through two given stops in Fig. 19 according to ‘Week’ and ‘24 Hours’, respectively; two routes, namely, 304 and 307, were used by the majority of elderly passengers, and the rush hours were during the time intervals 8 am to 10 am and 1 pm to 3 pm, for 304 and 307, respectively. These observations may provide clues to social welfare workers for analysing the travel behaviour of elderly people living in areas near these two routes.

4 Discussion and future work

4.1 Comparison with related studies

With the availability of electronic cards and related sensors, it is expected that passenger-trip records can be automatically collected, and different types of travel behaviour in public transportation can be analysed [13–15]. Unlike previous offline studies on bus transportation in Taichung city, which focused on computing and analysing the statistics of a specific route [16], a division [17], or the usage of specific cards [18], this study provides an interactive web interface to browse various class frequency distributions of any segments on all routes. That is, anyone can select arbitrary segments on a route, and subsequently inspect the class frequency distributions for different fixed-length time intervals (e.g. 24 h per day, day of the week, month and year).

To overcome the problems related to the computation of passenger-volume frequency, this study adopted the approach in [7, 8], whereby maximal repeats are extracted from tagged passenger trips, and various class frequency distributions are computed from a large number of passenger trips, with the class types being derived from the tags. The tags may consist of features from all passenger trips, such as two timestamps for embarkation or disembarkation, the types of electronic passenger cards, and the route number (identifier). It should be noted that as this approach is based on the Hadoop MapReduce programming model [9], it is scalable and has been successfully applied to real applications in other fields [19–22].

4.2 Providing more robust and precise class frequency distributions

The raw passenger-trip data in Table 2 corresponds to the period 2015âĂŞ2016. It is highly desirable to include more recent electronic records (i.e. since 2017) to improve the applicability of the web query system. Moreover, the processes of collecting the records of passenger trips by bus companies should be scheduled and pipelined automatically so that the quality of raw data can be improved. In this study, some records might have been missed when certain frequency distributions were inspected, but an empty slot was found. Furthermore, only two timestamps were recorded per trip, namely, embarkation and disembarkation. However, buses are currently equipped with GPS, so that the timestamps for all stops on a route can be recorded. Once these timestamps are available in ordered bus-stop lists, as in [19], it is possible to

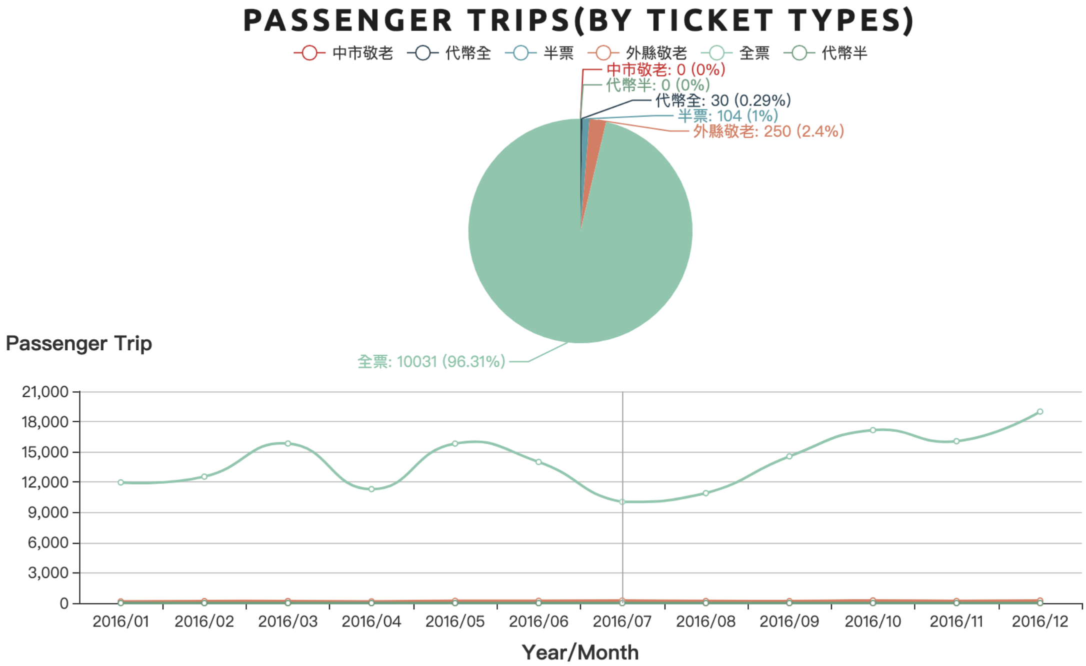

Fig. 15 *Monthly frequency distributions of passenger volume with conditions in Fig. 13 at month '2016/7'*

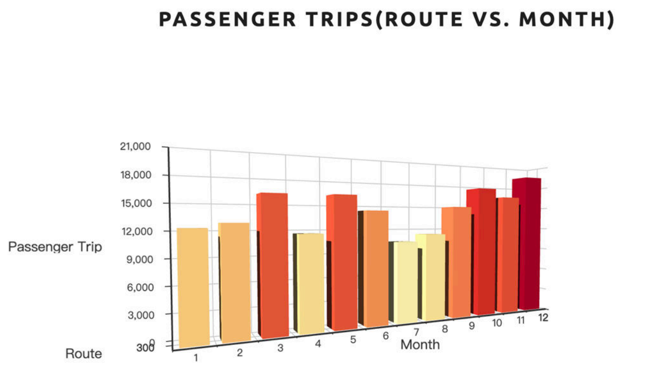

Fig. 16 *'Month' versus frequency distributions of passenger volume with parameters in Fig. 13*

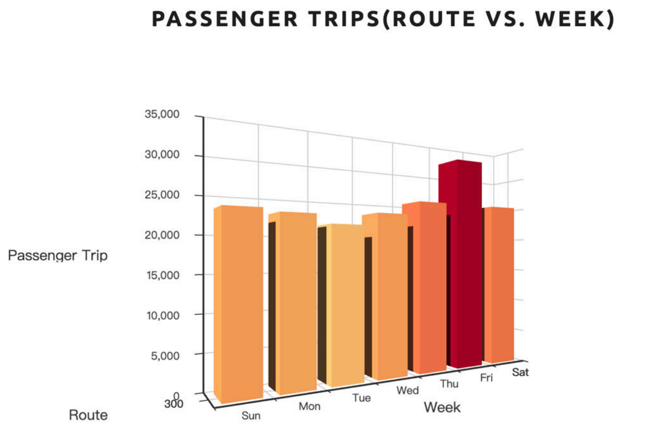

Fig. 17 *'Week' versus frequency distributions of passenger volume with parameters in Fig. 13*

provide more precise class frequency distributions using the scale 'minute', rather than 'hour'.

4.3 Migrating computation and storage to cloud computing platform

To meet the increased computational and storage requirements of the proposed system, a private Hadoop cluster with ten computing nodes (two masters and eight slaves) was used. The specifications of this cluster are given in Table 3. Nevertheless, ∼14 h and 50 min

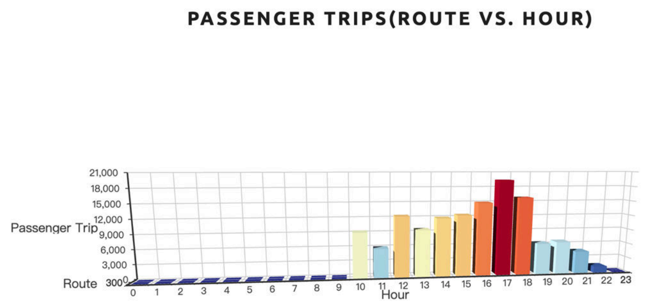

Fig. 18 *'24 Hours/ day' versus frequency distributions of passenger volume with parameters in Fig. 13*

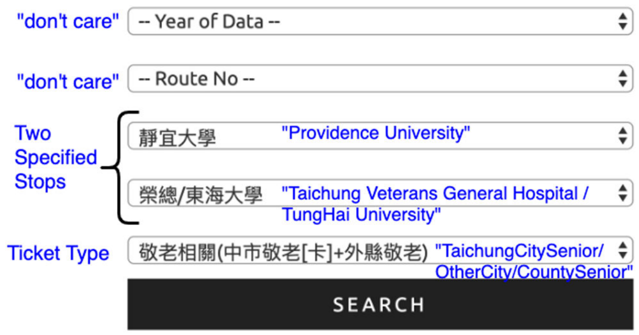

Fig. 19 *Example of query parameters for the elder residents in Taichung with two stops specified*

were required for extracting maximal repeats and computing various frequency distributions for these repeats. It is expected that computation time can be further reduced. Owing to the scalability characteristics of Hadoop MapReduce programming, which was used for maximal repeat extraction, existing cloud services, for instance, Amazon EMR [23], Microsoft Azure HDInsight [24], or Google Cloud Dataproc [25], may be used to increase the number of computing nodes and speed up computation if necessary. Moreover, to provide a more efficient web query system to the public, cloud services such as Auto Scaling-Amazon AWS [26] and Data Warehouse-Amazon Redshift [27] can be used for improved availability and scalability, rather than a single server, as shown in Fig. 6.

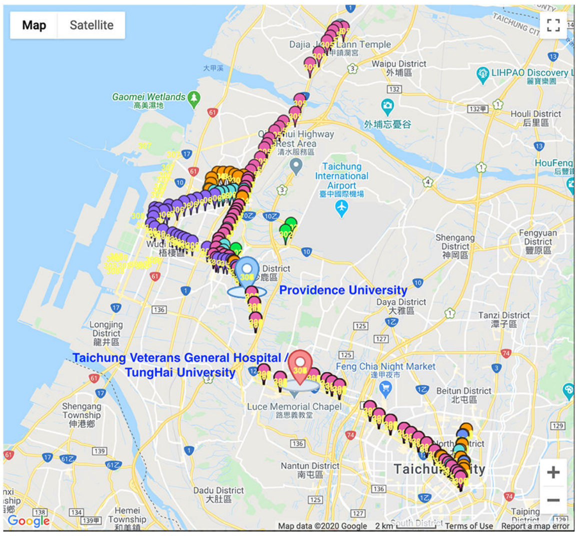

Fig. 20 *Google Map of all routes passing two specified stops in Fig. 19*

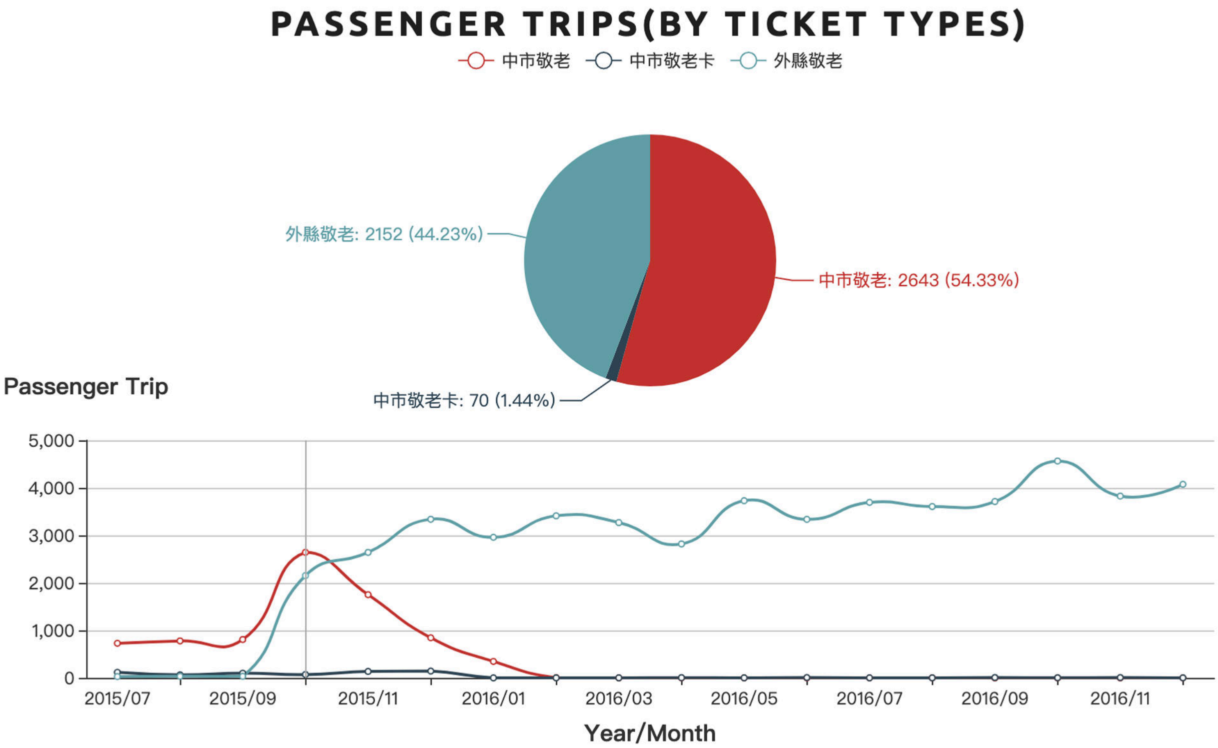

Fig. 21 *'Ticket Types' for 'Elder' versus frequency distributions of passenger volume when the month '2015/10' is pointed*

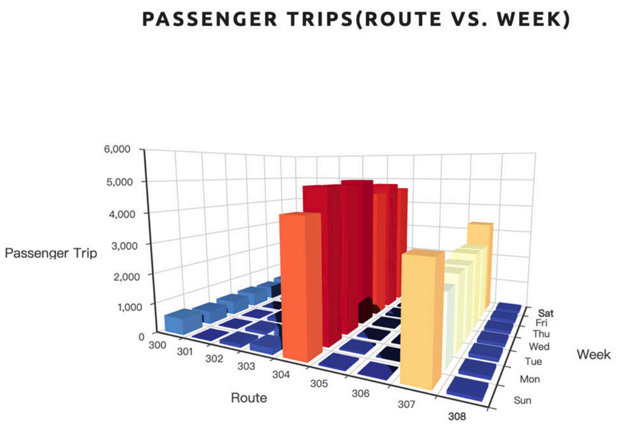

Fig. 22 *'Week' and 'Route' for 'Elder' versus frequency distributions of passenger volume*

4.4 Travel behaviour analysis for the elderly

As the problem of the ageing population in Taiwan is imminent, it is expected that better social welfare services would be provided if the government or domain experts could be able to capture patterns in the daily life of the elderly. In the study period, residents of Taichung city who were older than 65 could apply for an 'elder' card so that they could be eligible for discount fares in the public transportation system. These cards can link to personal information, such as an address, birthday, education and sex, for verification purposes. Therefore, social welfare experts may observe the frequency distributions for elderly passenger volume and capture daily-life patterns in this age group from a geographic point of view. Similarly, the local government can cooperate with bus companies to provide more suitable services and efficient schedules for these people. It would be interesting to undertake a behavioural analysis of the elderly using these cards if privacy issues could be resolved.

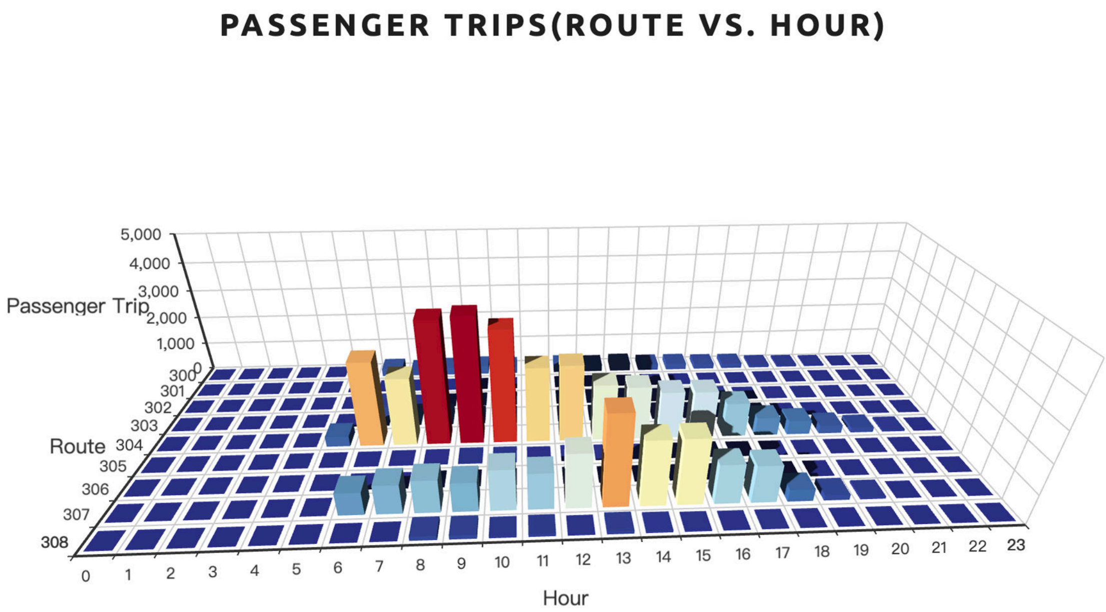

Fig. 23 *'24 Hours' and 'Route' for 'Elder' versus frequency distributions of passenger volume*

Table 3 Specification of private Hadoop (2 + 8) cluster

Types	Number of notes	CPU	RAM	Hard disk	OS	Hadoop cluster
master	2	(Intel Xeon Silver 4116 @ 2.10 GHz)*2	(32 GB DDR4 2400)*8	[2 TB*8], (SATA 7200 rpm) (RAID 6)	Centos 7.4	Cloudera Manger: CDH 5.12.2
slave	8	(Intel® Xeon® Processor E5-2630 2.30 GHz)*2	(16 GB DDR4 2133)*8	3 nodes: [2 TB*2] (SATA 7200 rpm) 5 nodes: [12 TB*2] (SATA 7200 rpm)	Centos 7.4	Hadoop: 2.6.0 + cdh5.14.2 + 2748

4.5 Service migration to other modern cities

As the number of residents is growing and traffic congestion is becoming a serious issue in modern cities, the efficiency of public transportation should be improved. If a city government establishes an electronic system to collect automatically all transactions related to passenger trips in the public transportation system, regular patterns in travel behaviour may be captured, and therefore transportation schedules or policies may be adjusted accordingly. This study provides not only methods for computing statistics related to passenger volume, but also suggestions for implementing a web query system. In future work, we will focus on migrating this system to other cities or countries.

4.6 Adjustment of local transportation policy

As the number of citizens is growing and traffic congestion problem is becoming serious in a modern city, the popularity and efficiency of public transportation are expected. That is, if one city government can establish an electronic system to collect automatically all of the transactions of passenger trips from the public transportation system, one can observe and find the regularities of passengers' behaviour and its impact on local government transportation policies, namely bus service, facilities and fare subsidy.

For bus services with regard to travel cost, walking and waiting times, seat availability, various socio-economic factors [28] and attitude and behaviour of driver [29], the government can adjust the policy or modify the schedule of public transportation in time. The most important service for the elderly is comfortability. Driver attitude and behaviour issue is the key point about the comfortability of elderly passenger when vehicles start to move or stop. The road operators must, therefore, give their drivers sufficient training and advice to raise awareness of driving behaviour and to evaluate the drivers' performance at the right time [29]. The facilities, such as priority seats, should be increased to meet future ageing trends; walking and waiting times of city bus should be shorter than before. However, above-mentioned policies are costly for the bus company and local government should provide extra subsidy from welfare funds in order to meet the cost.

5 Conclusion

This paper proposed a web query system with a visual interface that provides various frequency distributions of the passenger volume between any two stops on a bus route in Taichung city. These distributions were computed according to different fixed-length time intervals defined in advance, specifically, 'Year', 'Month', 'Week' and '24 Hours'. Furthermore, electronic-ticket types and routes could be combined for more precise observations. More importantly, the distributions did not refer to embarkation/disembarkation stops only, but to any two given stops.

The computation of the frequency distributions is scalable through a Hadoop cluster and can provide more robust statistics for passenger volume by combining tags (features) attached to passenger trips if necessary, rather than by simply analysing a dedicated route with limited stops, as in previous studies. Furthermore, a public web query system was developed (located at http://ttbd.asia.edu.tw/MRP/en/) and a guest account was set up. This system can provide various statistics for the passenger volume between two specified stops. For comparison, we used another web interface located at http://ttbd.asia.edu.tw/demo/en/, which considers two specified stops for embarkation and disembarkation.

It is desirable that the public can inspect various frequency distributions of passenger trips for any segments on a bus route. By providing statistical measures for passenger volume, the proposed system is expected to allow the global monitoring of the efficiency of a public transportation system through online services. Based on public responses in a smart city, bus companies can adjust their schedules, and the government may develop more effective public transportation policies for the elderly.

6 Acknowledgments

This research was funded by Asia University under project 107-asia-07 and project 10636073. The Taichung City Smart Transportation Big Data Database was used in this study and it was provided by the Bureau of Transportation, Taichung City Government. The authors would like to express our gratitude to the Taichung City Government and Asia University for their support. The authors thank Yong-Sheng Zhang for sharing the map of the bus route 300 – 310 in Taichung city.

7 References

[1] 'Public Transportation Information of Taichung'. Available at: https://tptis2015.blogspot.com/

[2] 'Public Transportation Information of Taichung for Taiwan Avenue'. Available at https://tptis2015.blogspot.com/2017/10/blog-post.html

[3] 'Easy Card'. Available at: http://www.easycard.com.tw/

[4] Wang, J.D., Lin, Y.N., Pan, S.H.: 'Analyzing the regularities of passengers according to different time intervals via local electronic bus system data in Taiwan'. The Institute for Operations Research and the Management Sciences, Taipei, Taiwan, 2018

[5] Wang, J.D., Pan, S.H.: 'A query system for cross-statistics of passenger trips within any segment of bus routes in taichung city - taking the records in 2015 and 2016 for example'. 2019 Conf. on Information Technology and Applications in Outlying Islands (ITAOI2019-IS16), Taichung, Taiwan, 2019

[6] 'ECHARTS'. Available at: https://www.echartsjs.com/en/index.html

[7] Wang, J.D.: 'Extracting significant pattern histories from timestamped texts using mapreduce', *J. Supercomput.*, 2016, **72**, (8), pp. 3236–3260

[8] Wang, C.T.: 'Method for extracting maximal repeat patterns and computing frequency distribution tables', Google Patents, 2019. US Patent No. 10,409,844

[9] Li, F., Ooi, B.C., Özsu, M.T., *et al.*: 'Distributed data management using mapreduce', *ACM Comput. Surv.*, 2014, **46**, (3), pp. 31:1–31:42

[10] Cormen, T.H., Leiserson, C.E., Rivest, R.L., *et al.*: '*Introduction to algorithms, third edition*' (MIT Press, Cambridge MA, 2009)

[11] Gusfield, D.: '*Algorithms on strings, trees, and sequences: computer science and computational biology*' (Cambridge University Press, New York, 1997)

[12] 'Hadoop'. Available at: https://hadoop.apache.org/

[13] Pelletier, M.P., Trépanier, M., Morency, C.: 'Smart card data use in public transit: A literature review', *Transp. Res. C, Emerg. Technol.*, 2011, **19**, (4), pp. 557–568

[14] Alsger, A.A., Mesbah, M., Ferreira, L., *et al.*: 'Use of smart card fare data to estimate public transport original-ddestination matrix', *Transp. Res. Rec.*, 2015, **2535**, (1), pp. 88–96

[15] Wang, W., Attanucci, J., Wilson, N.: 'Bus passenger origin-destination estimation and related analyses using automated data collection systems', *J. Public Transp.*, 2011, **14**, pp. 131–150

[16] Ho, C.Y., Chiu, I.H.: 'Research on passenger carrying capacity of taichung city bus with big data of electronic ticket transactions: A case study of route 151', *Commun. Comput. Inf. Sci.*, 2019, **1013**, pp. 238–249

[17] Ho, C.Y., Chiu, I.H.: 'The ridership analysis on inter-county/city service for the case study of taichung city bus system', *Commun. Comput. Inf. Sci.*, 2019, **1013**, pp. 250–259

[18] Ho, C.Y., Liao, S.C., Lien, Y.N., *et al.*: 'Applying the big data of electronic tickets to understand the behaviors of passengers with the senior cards and with non-senior cards in public transport âĂŞ A case study of taichung city bus'. Proc. of 2018 INFORMS Int. Conf., Taipei, Taiwan, 2018

[19] Wang, J.D., Hwang, M.C.: 'A novel approach to extract significant patterns of travel time intervals of vehicles from freeway gantry timestamp sequences', *Appl. Sci.*, 2017, **7**, (9), p. 878

[20] Wang, J.-D.: 'A novel approach to improve quality control by comparing the tagged sequences of product traceability'. MATEC Web Conf, Nantou, Taiwan, 2018, vol. 201, p. 05002

[21] Wang, J.D.: 'A novel approach to mine for genetic markers via comparing class frequency distributions of maximal repeats extracted from tagged whole genomic sequences'. In Abdurakhmonov, I.Y. (Ed): '*Bioinformatics in the era of post genomics and big data*' (IntechOpen, Rijeka, 2018), pp. 71–91

[22] Wang, J.D.: 'Reducing the gap between phenotypes and genotypes via comparing tagged whole genomic sequences'. The 12th Int. Conf. on Advancements in Bioinformatics and Drug Discovery (Journal of Proteomics & Bioinformatics), Dublin, Ireland, 2018

[23] 'Amazon EMR – Amazon Web Services'. Available at: https://aws.amazon.com/emr/

[24] 'Microsoft Azure HDInsight'. Available at: https://azure.microsoft.com/en-us/services/hdinsight/

[25] 'Google Cloud Platform- CLOUD DATAPROC'. Available at: https://cloud.google.com/dataproc/

[26] 'Auto Scaling – Amazon AWS'. Available at: https://aws.amazon.com/autoscaling/?nc1=h_ls

[27] 'Data Warehouse – Amazon Redshift'. Available at: https://aws.amazon.com/redshift/?nc1=h_ls

[28] Wong, R.C.P., Szeto, W., Yang, L., *et al.*: 'Public transport policy measures for improving elderly mobility', *Transp. Policy*, 2018, **63**, pp. 73–79

[29] Wong, R.C.P., Szeto, W., Yang, L., *et al.*: 'Elderly users's level of satisfaction with public transport services in a high-density and transit-oriented city', *J. Transp. Health.*, 2017, **7**, pp. 209–217

8 Appendix

8.1 Classification of ticket types

There are five types of electronic tickets in 'Taichung City Smart Transportation Big Data Database', which is provided by the Bureau of Transportation. The electronic ticket types and their owners' qualifications are as follows:

TaichungCitySenior: It may be owned by a person aged 65 or over who has established his/her household registration in Taichung and a Taiwanese aborigine aged 55 or over who has established his/her household registration in Taichung.

OtherCity/CountySenior: A senior card not issued by the Taichung City Government. It may be owned by a person/Taiwanese aborigine aged 65/55 or over, but who has established his/her household registration in another city/county, not in Taichung.

Half-fare: (i) Children between the ages of 6 and 12, (ii) elderly people over 65 years of age who do not have a Senior Card and (iii) people with disabilities and one of their companions are eligible.

Full-fare: A person who does not meet the above criteria should pay full fare.

Full-fare/Half-fareToken: This refers to a passenger who pays for a bus journey with cash, that is, without using an electronic ticket. In practice, when a passenger boards the bus, the driver will issue a token to the passenger. When the passenger wants to disembark, he/she should check the fare by tapping the token to the electronic ticket reader, then pay the fare in cash, and return the token to the driver.