A Bus-Arrival Time Prediction Model Based on Historical Traffic Patterns

Haitao Yu, Randong Xiao, Yong Du, Zhiying He
Beijing Transportation Information Center, Beijing, China 100161
Email: yuhaitao@bjjtw.gov.cn, xiaorandong@bjjtw.gov.cn

***Abstract*—The emphasis of this research effort was on using historical GPS data to develop a dynamic model to forecast long-term path travel times between bus stops of origin and destination. This study was brought about by the shortcomings of the existing real-time short-term prediction system, which have been developed and utilised in Beijing. The developed model is based on cluster analysis and polynomial fitting for the prediction of running times between bus stations. The research also intends to test the proposed model using real-world data. In a test with real-world bus data in Beijing, China, the proposed model performed effectively in terms of both prediction accuracy and computing time.**

***Keywords*—Bus-arrival time, prediction model, cluster analysis, polynomial fitting.**

I. Introduction

Deployment of intelligent transportation systems (ITS) is one of key areas of research and development in China's 12th five-year plan. Traffic congestion has been the serious problem facing Beijing for decades. Recently, Beijing's government has made tremendous efforts to solving the problem, and bus-arrival time prediction system will likely play an important role in the solutions. We own bus location along with other information such as time, occupancy, etc. Such information which has great potential as input data for kinds of applications including bus-arrival time prediction, is derived from 7193 vehicles equipped with GPS devices form 382 bus lines-131 bus routes operated by BFD BUS CO.,LTD and 25 bus routes operated by XL BUS CO.,LTD covering the entire Beijing.

Effective prediction of bus arrival time is central to our systems. To ensure the high accuracy of prediction, real-time road condition, which at least including vehicle velocity, traffic flow, degree of crowdedness and the state of traffic lights, is essential to prediction algorithm. Transit buses operations are usually disturbed by congestion on service route at different times of the day, intersection delays and excessive dwells at stops. The resulting impact of these factors comprises increasing passengers waiting times, increasing operation cost and traffic delays, etc. Thus, all of there factor will reduce the level of transportation and discourage passengers to use the buses. Therefore, one way to mitigate the impacts is to provide accurate information of bus arrival time at the stops. As for Beijing, our real-time system provide accurate real-time transit vehicle-arrival time information to the public, improving transit level of service. With accurate vehicle-arrival information, users may efficiently arrange their departure time from starting position and they can catch the buses or make successful transfer at transit stations, reducing passengers waiting times at the stops, and thus, enhance the quality of service.

TABLE I: Prediction result analysis of all lines

Distance	Prediction number	Average accuracy rate (ϵ)
all	19512686	25.25%
>700m	4384606	41.29%
<700m	15128080	20.60%

As illustrated in table I, prediction accuracy is shown different from distance derived from our real time system. The data were collected from 316 bus routes in the first week of November 2012 and percentage error is defined as follow.

$$\epsilon = \frac{Incorrect\ prediction\ number}{The\ total\ forecast\ number} \tag{1}$$

In order to improve the accuracy of prediction, we predict travel time according to historical data as a bus arrival time subsystem. The subsystem is used to handle long-term low-accuracy prediction. In this study, hierarchical method based on seuclidean is used to cluster historical data including running times and headway between stops, and weighted least squares method is adopted to establish the polynomial fitting function, indicating the dwell time and headway, and predicting the running time when the bus reaches the stopping stations. Traffic flow has the characteristics of periodic cycle for a week (i.e the velocity is generally consistent at the same time of a week). It can be observed that in all cases, there are five characteristic dates which are proposed based on the analysis of large amounts of historical vehicle data. The characteristic date refers to the dates with similar traffic flow in a week, five characteristic dates in this research consist of Sunday, Monday, Friday, Saturday and another characteristic date, which includes Tuesday, Wednesday and Thursday.

The rest of this paper is organized as follows. We review the related work in the next section. In Section III, we introduce our prediction model using the historical patterns including data clustering method, polynomial fitting based on the three prior stations and building of historical patterns. Section IV presents the numerical test using all 316 bus lines of Beijing, and specially take 944 route for a test bed. Finally, the conclusions are drawn in section V.

978-0-7695-5125-8/13 $26.00 © 2013 IEEE
DOI 10.1109/CSA.2013.87

IEEE computer society

II. Related work

Provision of accurate bus arrival information is crucial to passengers for reducing their anxieties and waiting times at bus stop. In recent years, many countries have came to realize that the accurate bus arrival time prediction play an important role in intelligent transportation system technology. By using advanced traffic information collection equipment, combined with geographical conditions and a variety of traffic impact factors, researches predict real-time bus arrival time, reducing the passengers' waiting time and improving the attraction of public transportation, and setting up the good image of the city finally.

In Japan, as one of transportation measures, the bus arrival time prediction system has got strongly support of government, obtaining the location of the bus through GPS devices and the control center, bus arrival time are send to user's computer or mobile phone. Experimental results show that the passengers' waiting time reduce 6 minutes (about 63% of the average waiting time), more than sixty percent of okayama residents regard as forecast system has make them more willing to take the bus, thus effectively limits the number of private cars [1].

In San Francisco, bus routes information are transmitted to the control center through Muni transport system, according to the actual location of the bus, arriving site and the typical traffic condition, estimated arrival time are transmitted to passengers over the electronic bus stops. A similar system is also used in Denver Rehoboth beach. Presently, there are two prediction algorithms widely used in the United States-algorithm designed for Virginia countryside Blaksburg based on GPS, and prediction algorithm put forward by the university of Washington in Seattle [2]. In China, vehicle traffic information service system based on city traffic information grid was developed by Tongji university, which realizes resource sharing and promotes computing efficiency with grid in selection of optimal travel scheme [3].

Previous studies have identified a range of determinants of bus travel time primarily via examining data from Advanced Public Transportation Systems (APTS) such as Automatic Passenger Counting (APC), Automatic Vehicle Location (AVL) and Global Position Systems (GPS) systems [4]. Existing methodologies include Regression models, Kalman filter models, Artificial Neural Network (ANN) models, and Support Vector Machines (SVM). In the work conducted by Patnaik et al. [5], the authors investigated the regression model using data collected by APC units installed in buses to estimate vehicle arrival times at all downstream stops. In another study [6], Chien et al. developed an ANN model to predict dynamic bus arrival time. In this research, two artificial neural networks (ANNs), trained by link-based and stop-based data, are applied to predict transit arrival times. An adaptive algorithm was applied to adjust the prediction based on the difference between the predicted and actual arrival times of the bus dispatched previously. However, the traffic data were sourced from simulation modelling. In [7], the authors presents support vector machines (SVM), a new neural network algorithm, to predict bus arrival time, examining the feasibility and applicability of SVM in vehicle travel time forecasting area. Dailey et al. developed an algorithm to predict bus-arrival time based on Kalman filter, which takes data collected by the on board automatic vehicle location system as input. The time and distance to the bus stop is predicted once latest information on bus location and time is obtained [8].

TABLE II: Line-link data set

Variable	Description
MapID	map idicator number
ID	line-link identity
Width	link width
Direction	the direction of the link
PathName	path name passing by the link
SnodeID	start node identification
EnodeID	end node identification

Most of the above model using real-time GPS data to predict bus arrival time, but when current position farther from the next station, because of the high time and spatial correlation of traffic flow, the average speed of the bus between the stops is not accurate (especially the complicated traffic condition in Beijing). In this paper, we first classified a week into five categories, and build a historical pattern based on cluster analysis. Historical data-based models predict travel time for a given period using the average velocity for the same time period obtained from a historical databases, each model indicates a period time of a characteristics day (half an hour in this research) as a supplementary subsystem. These models suppose traffic patterns are cyclical and the running time between the stops remain constant in the same characteristics day. We research and implement the subsystem in this work to solve the problem of long-term inaccurate prediction and other inaccuracies of real-time system.

III. Model development

A. Sampling point

GPS data of the buses are discrete point data available only at given time intervals instead of continuous point of data, the GPS devices have been installed in the buses running along each bus route to monitor bus operational information every 15 seconds or 20 seconds. In order to estimate when the bus would arrive at station B, information in addition to the bus location data must also include bus direction and the actual bus travel distance, instead of the Euclidean distance between station A and B. To obtain the spatial and temporal content of predicted bus, primary data process is done in three steps:

1) According to 2010 version of SURVEY map, the bus routes have been split by intersections, a line-link refers to the path between the intersections;

2) With the purpose of improving the matching accuracy, digitize the line-link into a set of nodes on a regular map spaced at 20 meters apart;

3) number every bus line and construct the serialization of the bus line-link.

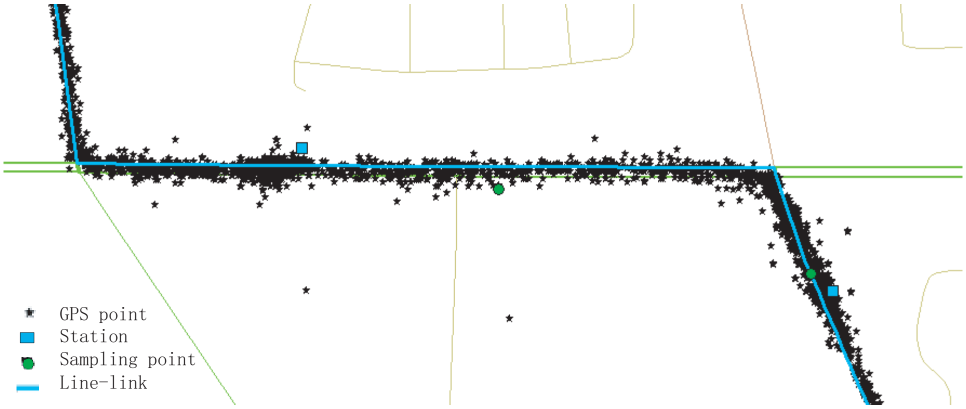

Fig. 1: Bus line link with GPS trajectory

As demonstrated in Table II, there are four different direction of the link, 0/1-bidirectional link, 2-along with the direction of the arrow, 3-contrary to the direction of the arrow.

B. Data

This study was based on data recorded start from 2011. The data contained a total of 382 lines, average spanning 40 intended bus stops in each direction. In general, each line serves more than thousands of passengers. Most of the buses has been installed with GPS units, The system tracks vehicles using an on-vehicle GPS-based vehicle computer which transmits vehicle position data to data center where data are processed and made available to the real time prediction system.

Fig.1 demonstrates a bus line link that was generated by the GPS trajectory received from bus every 15 seconds combined with Beijing digital map. The principle of digital line generation is based on GPS trajectory and then we modify, add, and remove the wrong line in return according to Beijing digital map.

C. Day-of-the-week velocity feature

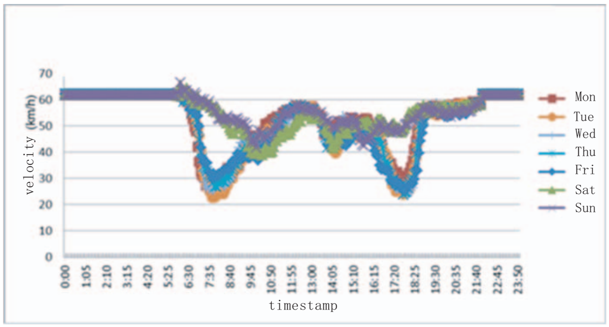

Fig. 2: Change tendency of average speed during a week

The data were collected from August 2012 to October 2012, bus transit path travel speed observations made approximately same at the same day of the week, they also show two bottom everyday corresponding to the morning peak and the evening peak. As demonstrated in Fig.2, on weekdays, two peaks occur each day, one in the morning and the other in the evening. A peak also occurs in the middle of the day on the weekend. It also shows that there are approximately five curve, and during Tue, Wed and Thu, there presents the same speed tendency. Traffic flow has the characteristics of periodic cycle for a week.

In order to prevent the effects of different patterns on prediction inaccuracy and to maximise the effects of similar patterns on prediction accuracy, a hierarchical cluster analysis technique is employed using Seuclidean distance to search for nearest period time in each day-of-the-week.

$$P(m) = [m,\ t_m;\ n,\ t_n] \tag{2}$$

Where

$P(m)$: Pattern set for day(m) of the week; and m, n: Elements of day-of-the-week set={Mon, TWT, Fri, Sat, Sun}, where TWT={Tue, Wed, Thu}. t_m: time period of the day(m) of the week, time period is divided the day into 48 hours, every half an hour is a time period, marked from 0 to 47.

D. Cluster analysis

By not considering historical average states, we classify the historical data into the time-of-day, day-of-week, as stated above. The input state is time-series dependent, in order to minimise uncertainties, we consider the dimension of time-series and cluster speed between continuous four stops at each time-of-day and day-of-week.

The clustering procedure is outlined as follows.

step 1: Using the average distance method, each data object is classified as a class, a total of N class. Set of data object is defined as:

$$DS = \{N_i | i = 0,\ 1,\ 2,\ 3\} \tag{3}$$

where $N_i = \{\nu,\ \ell\}$, ν is average arrival velocity, and ℓ is arrival distance.

step 2: Find the two nearest calss and merge its into one class, as result of the total class number decreasing one.

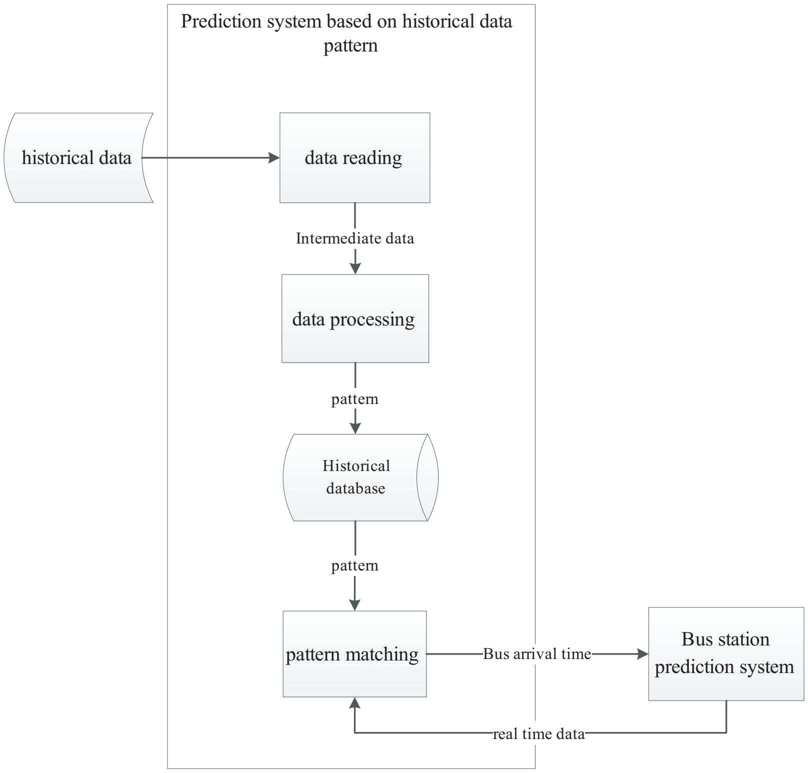

Fig. 3: Prediction model

The distance is defined as:

$$S_{ED} = \sqrt{\sum_{i=0}^{3} (\nu_{Ai} - \nu_{Bi})^2 w_i^2} \tag{4}$$

where $\nu_{Ai},\ \nu_{Bi}$ are the different velocity of the data class, and $w_i = \ell_i / \sum_{j=0}^{3} \ell_j$.

step 3: Calculate the distance between new class and all the old classes.

step 4: Repeat steps 2 and step3, until the distance between all the classes is greater than the threshold, the threshold value is defined as $T = 2.77 \times \sqrt{\sum_{i=0}^{3} w_i^2}$.

E. Polynomial fitting

After clustering analysis, as illustrated in Fig.3, the data processing module includes the data clustering and polynomial fitting. The polynomial fitting function used in this work:

$$\widehat{v} = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4$$

The least squares estimate method is used to calculate the coefficient vector $\widehat{a} = [a_0,\ a_1,\ a_2,\ a_3,\ a_4]$. There output of fitting are insert into the database of patterns in the form of $(time\ period,\ line,\ station\ NO,\ \widehat{a})$. There are five database of patterns correspond to the five day-of-the-week {Mon, TWT, Fri, Sat, Sun}.

F. Pattern matching and prediction

In the process of prediction, the velocities of three prior stops are used to predict the next stop velocity,the message (day-of-the-week, time period, line, station NO h, the third stop running time, the second stop running time, the first stop running time) are changed to (day-of-the-week, time period, line, station NO h, the third stop speed, the second stop speed, the first stop speed) as the input vector. Suppose the minimal station number is k, a special case is that $h-k < 3$, in another work the station number is not enough. In this case, station NO is set to $k+3$. Then, according to the prediction system, input vector is used to match the nearest historical pattern.

The pattern matching procedure is outlined as follow.

step 1: Query the historical patterns using the current day-of-the-week and period time, obtain the set

$$P = \{p_i,\ |i = 0,\ 1,\ 2, \cdots,\ k\}$$

step 2: For each element of historical data, calculate the distance S_{ED_i} between the current vector and all patterns obtained above using the Equation (4).

$$S_{ED_{min}} = min\{S_{ED_1}, \cdots,\ S_{ED_k}\}$$

step 3: Estimate $t = \ell_{min}/\hat{v}$.

IV. NUMERICAL TEST

The interval prediction method is based on historical patterns, as described in above section. In this section, the proposed model is tested and analysed in order to demonstrate the accuracy and efficiency. In order to provide a comparative study, a performance measure is applied to evaluate the variation between the predictions and the observations.

A collection of the buses on this line were equipped with GPS devices recording bus arrival times corresponding to each stops. A three month travel velocity data set (starting from August 2012) was supplied for the study. This data was required because historical patterns are essentially based on vast quantities of historical data [9] [10]. Considering the data aggregation scheme and the time horizon of the prediction model, continuous time was divided into discrete time based on the period time and day-of-the-week.

In order to test and verify the validity of the model, the 944 bus line in Beijing was used as a test bed, because most of the site is far apart. The in-vehicle GPS receiver measures the location and time of buses for every second and transmit measured data to the data center at 15 second time cycle. The 944 line is made up of 61 stops between NAIZIFANG and CAIHUYINGQIAOXI in one direction and 62 stops in another direction, we test in the direction of 61 stations. The length of the route is 46.1km, and the bus runs start from 5:30 am to 21:30 passing through the north and the south of Beijing.

TABLE III: Prediction result analysis of 944

Distance	Station number	Prediction number	MAPE
all	61	29269	21.41%
<700m	27	13634	28.86%
>=700m	34	15635	14.91%

TABLE IV: Prediction result analysis of all lines

Distance	Station number	Prediction number	MAPE
all	10629	2313619	22.57%
<700m	5071	1101789	29.47%
>=700m	5558	1211830	16.29%

The performance of the developed model for prediction accuracy is illustrated in Tab III and Tab IV through comparative analysis between the long and short distance differentiate by 700 meter. The Mean Absolute Percentage Error (MAPE) is used to evaluate the performance, it provides the most useful basis for comparison [9]. MAPE is defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{y_i} \times 100 \qquad (5)$$

where

n: Sample size
y_i: Actual running time
$\hat{y}_i$: Prediction running time

V. Conclusion and prospect

In this paper, firstly, we find that bus velocity has time correction through vast real-world traveling speed of 316 bus lines. Real-world historical GPS data are used in developing travel-time prediction methodology, which is based on cluster analysis and polynomial fitting for the prediction of running times between bus stations. The hierarchical clustering method and seuclidean distance method are chosen because of different similarity of data set and lowest average error rate in applying real-world test. Prediction models are developed under different patterns. Given the bus line, period time and day-of-the-week, the method shall generate historical patterns and build estimated travel velocity on the next stop along the route based on prior three stations.

The model was tested with a real-world data and performed effectively prediction accuracy in the dimension of both strategy for providing travel time information and real-time application as a subsystem. It is considered that the developed model is simple enough to avoid the need for extensive computation time to access to the real-time system. Since our model does not consider the effect of traffic light and traffic congest, which is crucial to the prediction accuracy, the accuracy remains to be further improved.

Acknowledgment

This work was supported by National High-tech R&D Program of China (863 Program) (NO. 2012AA12A207) and the Fundamental Research Funds for the Central Universities.

References

[1] H. Makino, “Smartway project,” *Development*, vol. 2005, 2005.

[2] S. Giering, *Public Participation Strategies for Transit: A Synthesis of Transit Practice*. Transportation Research Board, 2011, vol. 89.

[3] L. J.-x. Z. Hui-juan, “Model and analysis for transportation information grid system based on colored petri net,” *Microcomputer Information*, vol. 6, p. 011, 2011.

[4] E. Mazloumi, G. Currie, and G. Rose, “Using traffic flow data to predict bus travel time variability through an enhanced artificial neural network,” in *World Congress on Transport Research, 12th, 2010, Lisbon, Portugal*, no. 03377, 2010.

[5] J. Patnaik, S. Chien, A. Bladikas *et al.*, “Estimation of bus arrival times using apc data,” *Journal of public transportation*, vol. 7, no. 1, pp. 1–20, 2004.

[6] S. I.-J. Chien, Y. Ding, and C. Wei, “Dynamic bus arrival time prediction with artificial neural networks,” *Journal of Transportation Engineering*, vol. 128, no. 5, pp. 429–438, 2002.

[7] Y. Bin, Y. Zhongzhen, and Y. Baozhen, “Bus arrival time prediction using support vector machines,” *Journal of Intelligent Transportation Systems*, vol. 10, no. 4, pp. 151–158, 2006.

[8] D. Dailey, S. Maclean, F. Cathey, and Z. Wall, “Transit vehicle arrival prediction: Algorithm and large-scale implementation,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1771, no. 1, pp. 46–51, 2001.

[9] H. Chang, D. Park, S. Lee, H. Lee, and S. Baek, “Dynamic multi-interval bus travel time prediction using bus transit data,” *Transportmetrica*, vol. 6, no. 1, pp. 19–38, 2010.

[10] J. Biagioni, T. Gerlich, T. Merrifield, and J. Eriksson, “Easytracker: automatic transit tracking, mapping, and arrival time prediction using smartphones,” in *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems*. ACM, 2011, pp. 68–81.