Model of Delay Prediction for Signalized Intersection Based on GPS Data

Ya WANG$^{1,2,*}$, Chang-jun JIANG$^1$ and Hong-lei REN$^3$

$^1$State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
$^2$Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Nanjing, China
$^3$School of Software, Beihang University, Beijing, China

*Corresponding author

Keywords: Bus arrival time, Prediction model, Signalized intersection.

Abstract. To facilitate travel, many domestic bus transits have been fit with station prompt means. With this function, the bus arrival time can be forecast in real time based on the GPS data. However, the current bus arrival time prediction model features high assumption, not well adapted to urban roadway traffic congestion and delay at signalized intersection, thus compromising the accuracy of the prediction. A real time delay estimation for signalized intersection is proposed in this paper, combining the physical model derivation and data fitting, integrating the waiting time at signalized intersections into the existing prediction model. It is shown in the experiments that compared with previous models, the accuracy of our model increases by 29.1%. The mass GPS data is processed via the Hadoop platform, the performance is improved by 3600%~4000% compared with the serial program.

Introduction

In urban transits, the transit vehicle occupies a key position, which is arteries for urban production and living, and an important indicator of the level of urban modernization. However, the amount of private cars in our country has increased in recent years, not only resulting in road congestion, reducing the efficiency of the public transits, but causing unnecessary wasting time; In addition, the increasing emission causes increasing pollution. Therefore, to provide the public with an efficient, intelligent and comfortable bus travel is an urgent traffic problem at current stage of traffic development [1].

A good intelligent transit system must be able to reflect the real time state of transit vehicle on the road, and response to unexpected situations reasonably and timely via monitoring. Accordingly, the intelligent bus transit with bus arrival time prediction can provide better service to the public.

However, in the existing bus arrival time prediction model, the dramatic change to the average travel speed within short time is not taken into consideration, and prediction of the system at signalized intersection cannot achieve the desired results. In the past decades, the delay at signalized intersections and others are deemed as an important index for intersection traffic capacity by most traffic engineering experts. Most models for delay prediction at intersections are committed to describing the statistical characteristic of the overall delay, such as delay time distribution and the average delay time[2]. Many of the models are still at the stage of off-line analysis due to limit on acquisition of intersection traffic data [3]. In this paper, the severe fluctuation to the average road speed following switching of light lamps at intersection is studied, the vehicle delay caused by traffic lights are forecast in real time, providing an accurate vehicle stop delay prediction model, so as to obtain accurate bus arrival time via the bus arrival time prediction model. In addition, a data processing system based on Hadoop platform for parallel processing of these mass data is proposed in this paper.

The main contribution of this paper: 1) propose a real time delay at signalized intersection prediction model. It is verified in the experimental that, the accuracy of arrival time increases by 29.1% compared with existing models; 2) a data processing system based on Hadoop platform is proposed. The GPS data for buses in Beijing in one day can be handled within 15 to 20 minutes.
Bus Arrival Time Prediction Model

The arrival time prediction is to determine the locations of all buses using the stops/links data and GPS data, calculate the normal travel time from this location to the destination and the delay using the calculated travel time for each link, and then calculate the sum to get the arrival times of all operating buses to destinations.

Floating Car Data technology is used for the arrival time calculation based on real time traffic [4], the average speed of each link within short time is obtained first, then the travel time is the ratio of the distance to the speed. Then the travel times for the vehicle for all links are added to predict the bus arrival time. Take \( TravelTime \) as the travel time of the bus to the next stop, which is expressed as:

\[
TravelTime = \sum_i \frac{Length_i}{Speed_i}
\]  

where \( Length_i \) means the length of the previous \( i^{th} \) link; \( Speed_i \) means the travel speed for the previous \( i^{th} \) link; for the bus arrival time prediction system, the accuracy of speed of link will directly affect the accuracy of arrival time.

In the real time prediction model, we use the raw bus GPS data to calculate the link speed. The calculation of link speed is divided into two parts: lateral vehicle speed and longitudinal link speed. The lateral vehicle speed is defined as the average vehicle speed within the past 3 minutes. The time difference between the two GPS points for calculation time \( \text{Diff} = T_e - T_s \), where \( T_e \) is the time of the last GPS point, \( T_s \) is the time of the first GPS point, the traveled distance is \( \text{dis} \). The lateral vehicle speed \( \text{speedHor} \) is expressed as:

\[
\text{speedHor} = \frac{\text{timeDiff}}{\text{dis}}
\]

The longitudinal link speed is defined as the average speed of all vehicles with time difference within 30 minutes of recent 10 vehicles along a link, as shown in Figure 1: in the longitudinal link speed queue of link 4, there are 4 vehicles with time difference within 30 minutes of the recent 10 vehicles.

![Figure 1. Longitudinal link speed.](image)

Average velocity of Vehicle 4 on Link4: V4
Average velocity of Vehicle 3 on Link4: V3
Average velocity of Vehicle 2 on Link4: V2
Average velocity of Vehicle 1 on Link4: V1

After calculating the current lateral vehicle speed using the definition of the lateral vehicle speed, add the lateral speed of this vehicle to the longitudinal link speed queue so as to calculate the longitudinal link speed. Calculate the weighted mean of the longitudinal link speed depending on the weighted value. Take \( w_i \) as the weighted speed of the \( i^{th} \) vehicle, \( v_i \) is the average speed of the \( i^{th} \) vehicle along the link, the longitudinal link speed \( \text{speedLor} \) is expressed as:

\[
\text{speedLor} = \sum_i w_i \cdot v_i
\]

We can see that, the reasonableness of the above algorithm must be based on at least the two following assumptions: 1)The average speed of adjacent links within short time (3 minutes) is Table 2. The average speed of a link within short time (3 minutes) changes continuously.

We can see from the above analysis that, the road conditions considered in the bus arrival time prediction system based on the real time traffic are too simple, the requirement for the road conditions are high, which applies only to continuous congestion-free road conditions. Due to the normal congestion on urban roads, the congestion caused by signalized intersections becomes increasingly evident [5]. So the above assumption is difficult to hold, and will seriously affect the prediction.
Prediction Model for Delay at Signalized Intersections

For practical assumption of road condition in above models, we proposed a prediction model for delay at signalized intersections, so as to obtain more accurate forecast arrival time. For this, we need to discuss the traffic laws for vehicles passing signalized intersections.

To do this, we discuss the mode for vehicles passing signalized intersections from history GPS data. The picture shown in Figure 2(a) is the GPS track for a Bafangda Link 944 bus passing through Anyuan intersection, the sampling interval between GPS points is 20 seconds. There is a certain distance between two adjacent GPS points, indicating that the bus was travelling as it passed through Anyuan Road, and not affected by traffic lights. In Figure 2(b) and Figure 2(c), the bus stopped for one time and twice respectively. The GPS points in the red circle gathered, indicating that the bus stopped or was traveling at low speed, and the bus encountered red traffic light at the intersection.

Figure 2. (a) GPS track for the vehicle with 0 stops at the intersection; (b) GPS track for the vehicle with 1 stop at the intersection; (c) GPS track for the vehicle with 2 stops at the intersection.

It is known from the above analysis, as the bus stops and waits at the traffic lights, the GPS points gather, and the density of GPS points is high. By examining these gather, we can find out the number of stops and states of the bus at the intersection, which we are really interested. In particular, the identification of gathering of the data can be classified as clustering issue. Here we adopt the density-based DBSCAN clustering algorithm [6] to solve this problem.

Derivation of Model for Delay at Signalized Intersections

In this paper, a prediction model for delay at signalized intersection based on the assumption of proven steady delay model is proposed [7]. Based on this model, we studied and analyzed the stop and delay of a certain bus to determine the expression.

Based on the steady delay model, we propose the following assumptions for this model: 1) The signal timing is fixed, and the length of initial vehicle queue is 0; 2) the vehicle arrived at the intersection at even speed; 3) the average vehicle arrival rate within the stipulated period is stable (that is, the traffic flow is stable within the stipulated period); 4) The travel capacity at the intersection is constant within the study period, and the arrival rate cannot exceed the travel capacity; 5) Within the certain period T, the vehicle arrival rate changes randomly within each signal cycle, so the arrival and departure of vehicles may be imbalanced within some signal cycles, resulting in excessive queue vehicles (multi waiting vehicles); but after a number of cycles, the excessive queue vehicles will disappear. For the whole period T, the arrival and departure of vehicles is balanced.

Based on the above assumption, according to the derivation of modeling for vehicles passing through the intersection by Li et al. [6], we proposed a model to get the time for the vehicle to pass through the traffic light, the expression is:

\[
\text{StopTime} = t_0 - D_0 x \tag{4}
\]

where \(t_0\) is the signal timing for red lamp. By assumption (1), the signal timing is fixed, so for certain traffic light, this parameter is a constant; \(D_0\) is the traffic density. By assumption (3), the traffic density is not changed and known within a certain time. Within a certain period, \(D_0\) is a constant, from the traffic flow theory, this time is generally 30-60min[8]. The parameter \(x\) is the distance between the vehicle stop location to the traffic light. It can be seen that, the vehicle delay is proportional to the distance between the vehicle stop location to the traffic light. The parameter \(t_0\) and \(D_0\) are unknown constant parameters for specific traffic light and specific period, and the
parameter \( x \) can be determined as the vehicle stops. Therefore, to get the vehicle delay time, just get the unknown parameters \( t_0 \) and \( D_0 \) for different traffic lights and periods. In the following context, we will get the parameter \( t_0 \) and \( D_0 \) using the Hadoop platform.

**Bus Arrival Time Prediction Model Considering Signalized Intersection**

In order to take advantages of the real time model [9] and the prediction model for delay at signalized intersections, we classify the traveled link into the straight sections and range of traffic light. We use different models to predict the time in different sections, and add the times of all sections.

![Figure 3](image)

Figure 3. (a) Vehicle arrival time prediction model considering the traffic lights (b) Prediction model as the vehicle is within the range of traffic light.

As shown in Figure 3(a), the vehicle A is at the current location C, and there is a traffic light before the next stop. For each traffic light, we got the maximum queue length \( Q_{length} \) for this traffic light from the history stop data. Classify section Q as the range of traffic lights, and other continuous link as a single link for calculation. Set the average speed of section 1 and 2 as \( v_1 \) and \( v_2 \) respectively, and the history average time passing section Q to \( t_{avg} \) (the calculation method is described in the following section), then the time \( t \) for vehicle A arriving at the next stop is expressed as:

\[
t = t_{avg} + \frac{S_1}{v_1} + \frac{S_2}{v_2}
\]  

(5)

In addition to the case shown in Figure 3(a), there is another case to be considered, that is, vehicle A is traveling within the range of traffic light. As shown in Figure 3(b), vehicle A has left the range of traffic light, and the travel distance is \( Q_a \). Now there are two cases: 1) the vehicle doesn’t stop at red lamp. The expected time for the vehicle at the traffic light is the history average passing time \( t_{avg} \) for section Q subtracting the past time \( t_a \). In application, we can use the instantaneous speed from the bus GPS data to calculate the value \( t_a \), because the time for the bus to travel through section \( Q_a \) may be very short. In this case, the time \( t \) for vehicle A arriving at next stop is expressed as:

\[
t = t_{avg} - \frac{Q_a}{v_{inst}} + \frac{S_2}{v_2}
\]  

(6)

The vehicle stops to wait at red lights. As the vehicle stops to wait, we can calculate the current vehicle’s queue distance based on the location of the bus GPS points and the location of traffic lights in the map data, and then predicate the stop delay \( t_{delay} \) using \( t_0 \) and \( D_0 \) obtained via the regression model in expression in last section, then obtain the starting speed \( v_b \) after the vehicle stops to wait using history data, and calculate the prediction time for the vehicle to pass through \( Q_b \). At last, the time \( t \) for vehicle A to arrive at the next stop is expressed as:

\[
t = t_{avg} + t_{delay} + \frac{Q_b}{v_b} + \frac{S_2}{v_2}
\]  

(7)

The value \( t_{avg} \) is calculated as following. First, the range of Q will generally relatively small, usually 100-200m, and Q is the range of traffic light, the speed changes without law, so it is not reasonable to calculate the pass time as the ratio of the distance to the speed. The speed changes without law, so it is not reasonable to calculate the pass time via the ratio of the distance to the speed.

Obtain the time stamp for the vehicle to travel through the corresponding location using the proportional distances. As shown in Figure 4, suppose we will calculate the \( t_{avg} \) for the vehicle to
travel from point A to point B, and it is known that $t_1$, $t_2$, $t_3$ and $t_4$ are the time stamps for the bus to travel through adjacent GPS points respectively. Suppose the vehicle travels at uniform speed from $t_1$ to $t_2$ and $t_3$ to $t_4$. The value $S_1$, $S_2$, $S_3$ and $S_4$ can be calculated via the GPS location information. By assumption of uniform motion, we can calculate $t_A$ and $t_B$, represented by the following formula:

$$t_A = t_1 + \frac{S_2}{S_1} \cdot (t_2 - t_1)$$ (8)

$$t_B = t_4 - \frac{S_4}{S_3} \cdot (t_4 - t_3)$$ (9)

So the time for the vehicle to travel from point A to point B is $t_B - t_A$. Calculate $t_B - t_A$ based on the history data for multi times, then get the average to obtain $t_{avg}$.

$$t_{avg} = \frac{1}{n} \sum_{i=1}^{n} (t_{Bi} - t_{Ai})$$ (10)

In the traffic light model, point B is the location of traffic lights, and point A is the location of vehicle. Maybe there is no GPS data for the vehicle at point A and point B, So after correcting the time stamp for the vehicle to travel through the corresponding location using the proportional distances, the time difference between two time stamps is the interval, the value $t_{avg}$ from point A to point B is the average history time to travel through the range of traffic lights Q as described in the above model.

**Implementation of the Model Using Hadoop Platform**

In this section, we use the above-mentioned MapReduce paradigm to implement prediction model for delay at signalized intersections to adapt cloud computing environment, so that it can be implemented using Hadoop platform, to calculate a large number of vehicle GPS data, increase the amount of data in the training pattern library, to improve the training accuracy for the parameters of the model.

During implementation, we created a two-step MapReduce model. The first MapReduce process is performed to preprocess the history GPS data, checking all GPS data to filter "dirty" data, and then match the GPS data to the corresponding traffic light. The input data set of MR I is unordered raw GPS dataset, containing a lot of useless data. After reading the GPS data, Map function will extract the latitude and longitude of each GPS data, and will match the vehicle location of each GPS data to the link in the map using the map matching algorithm proposed by Qie et al [10]. Based on the data of traffic light layer, extract the latitude and longitude matching the traffic light, and calculate the distance from the GPS point to the traffic light. At last, the Map function will match the data within the range of traffic light to the traffic light.

At Reduce stage, DBSCAN clustering algorithm is adopted to cluster these GPS data to determine the stop state of each bus at current time. As above-mentioned, in order to ensure the amount of data in the training data set of model parameters, we need to save 90-day stop data in HBase database, so the RowKey is: TrafficLightID_TimeStap_carId. Such RowKey design will increase HBase data retrieval speed, which will be described in introduction to MR II.

In MR II, the Map function will read data from HBase, and merge by traffic light ID. In MR II, the RowKey is designed as TrafficLightID_TimeStap_carId. With such design, the data of the same
traffic light may be distributed centrally to the greatest extent, to reduce the loss of network transmission. After the shuffle procedure, the pairs are sorted by time and ID. The stopInfo for the same traffic light ID is automatically sorted using Hadoop.

**Results**

In order to verify the accuracy of the prediction model for delay at signalized intersections based on GPS data, the data for at signalized intersections is tested in this paper.

To test the accuracy of stop delay prediction in the traffic delay prediction system, we take history GPS data for 3000 buses in Beijing Bafangda Bus Co. Ltd, Xianglong Bus Co. Ltd and Bus Group for 72 days from April 27, 2015 to July 6, 2015. For real time data, take the Beijing Xueyuanqiao intersection data of July 13, 2015 8:00-9:00, 11:00-12:00 and 18:00-19:00.

The test results are shown in Figure 5 and Figure 6. In the figure, Abscissa is the deviation of the forecast time and the true time. The main ordinate represents the number of prediction. The secondary ordinate represents the cumulative percentage of number of prediction, that is, the percentage of the number of prediction with deviation within the range accounting for the total number of prediction. As shown in Figure 5, for the 8:00-9:00 period, the prediction deviation for Link 1 bus travelling through the intersection: the predictions with deviation within 10 seconds reaches 400 times, and the accumulative percentage accounts for nearly 40%; the predictions with deviation within 20 seconds accounts for 64%, the predictions with deviation within 30 seconds accounts for 79%. It is shown in Figure 6, within the same period, the number of prediction with deviation within 30 seconds for Link 99 bus also accounts for 79%. It can be seen from the above analysis, the overall accuracy of prediction system for delay at signalized intersections based on the GPS data is high, reaching the desired standard. The number of predictions with deviation within 10 seconds accounts for the most, and the number of predictions with deviation within 30 seconds accounts for less.

Currently in the prediction system for delay at signalized intersections, the deviation between the forecast time and the true time is within 20 seconds. It is deemed that this prediction result is accurate. We define the accuracy of prediction: The deviation between the real delay and the forecast time for a specific prediction. The calculation expression:

\[
C_i = \begin{cases} 
100\% & (|I_{nr} - I_{nf}| < 20s) \\
\frac{|I_{nr} - I_{nf}|}{I_{nr}} \times 100\% & (20s < |I_{nr} - I_{nf}| < I_{nr}) \\
0 & (|I_{nr} - I_{nf}| > I_{nr})
\end{cases}
\]  

where \(C_i\) is the accuracy of the \(i^{th}\) prediction; \(I_{nr}\) is the real delay; \(I_{nf}\) is the predicted delay.

The overall accuracy of the prediction of the data was calculated using the above formulas, and the result was 86%. This shows that the forecast value obtained using the prediction system is close to the actual value. Most of the predictions recorded by GPS are accurate, and the precision model algorithm has been verified.

![Figure 5. Stat. of Link 1 bus intersection delay prediction at 8:00-9:00.](image-url)
In order to analyze the improvement of bus arrival time prediction model by the delay at signalized intersections, we compared the traditional bus arrival model with the arrival prediction model integrating prediction for the delay at signalized intersections. The overall 7-day accuracy and change to the accuracy for the system is shown in Table 1. The deviation between the actual delay and the forecast times.

<table>
<thead>
<tr>
<th>Name of prediction system</th>
<th>7-day precision</th>
<th>7-day average accuracy</th>
<th>Max. absolute deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>New arrival time prediction system</td>
<td>82.16%</td>
<td>89.20%</td>
<td>110s</td>
</tr>
<tr>
<td>Old Beijing arrival time prediction system</td>
<td>80.52%</td>
<td>60.12%</td>
<td>254s</td>
</tr>
</tbody>
</table>

It is shown that after integrating of the models, many prediction deviation of the new system reduces below 30 seconds, and the max absolute deviation reduces from 254s to 110s.

**Conclusion**

The current bus arrival time prediction model features high assumption, not well adapted to urban roadway traffic congestion and delay at signalized intersection, thus compromising the accuracy of the prediction. A real time delay estimation for signalized intersection is proposed in this paper, combining the physical model derivation and data fitting, integrating the waiting time at signalized intersections into the existing prediction model. It is shown in the experiments that, compared with previous models, the accuracy of our model increases by 29.1%. The mass GPS data is processed via the Hadoop platform. It takes the serial program 10~12h to handle the Beijing GPS data for one day, and only 15~20min for the Hadoop platform, the performance is improved by 3600%~4000% compared with the serial program.

**References**


