Providing Useful Information for Passengers Based on TTF Model

Yangdong Liu\textsuperscript{1,2}, Ye Tian\textsuperscript{1}, Bo Yuan\textsuperscript{1}, Chang Wu\textsuperscript{1,2}, Weishuo Qian\textsuperscript{1,2}, Wei Mao\textsuperscript{1}

\textsuperscript{1}Computer Network Information Center, Chinese Academy of Sciences
\textsuperscript{2}University of Chinese Academy of Sciences

No. 4, South 4th Street, Zhongguancun, Haidian district, Beijing 100190, China
{liuyangdong, tianye, qianweishuo, mao} @cnnic.cn, wuchang12@mails.ucas.ac.cn

Abstract—Taxis are an important part of public transportation in nowadays. Taking a vacant taxi has become a difficult problem due to the imbalance between demand and supply of taxi service. People want to kno where is better to wait for a vacant taxi in order to save their time in the fast-paced city life. Helpful information could be provided to the passengers by mining the historical GPS traces generated by taxis. In this paper, we propose a model based on taxi traffic flow (TTF) model to measure the difficulty of taking a vacant taxi on the road segments nearby. Furthermore, we calculate probability and expected waiting time of taking a vacant taxi based on TTF model. Afterwards, we detect the hotspots on the road segments expecting to offer more available information. Our research is based on the real historical taxi GPS data generated by 12000 taxis in Beijing. We conduct experiments to validate our model eventually and the experimental results are in line with the actual situation well relatively.

Keywords— taxi traffic flow; waiting time; hotspots; taxi GPS data

I. INTRODUCTION

People have become accustomed to take taxis in their daily life. As the disequilibrium between the increasing population and limited traffic facilities is becoming more severe, we usually have to wait for a long time to take a vacant taxi especially in big cities, such as Beijing, Taipei and New York. There are so many factors that influence the difficulty of taking a vacant taxi, such as the location where the passenger is standing by, the time when the passenger wants to take a vacant taxi and weather condition. Passengers hope to get the real-time traffic information that helps them improve their experience of taking a vacant taxi.

The taxis equipped with GPS devices generate massive real-time trace data. For example, 50,000 taxis carrying 3 million passengers generate 60 millions of records every day. These records include the position, speed, status of the taxi that can reflect the taxi’s real behavior. The rich spatio-temporal information has attracted a number of researchers’ attention. A lot of studies have been done to investigate the useful information for taxi drivers or passengers [1, 2, 3, 4, 5, 13, and 14]. However, the existing work is not enough.

In this paper, we propose a tactics to help the passengers take a vacant taxi by analyzing a huge number of historical taxi GPS data. Given the current time and geo-position of a passenger, we measure the difficult of taking a vacant taxi on each side of a road segment nearby and then give some advices based on the real-time traffic information to the person expecting to take a vacant taxi. The information can help the passengers to choose a better road segment to wait for a vacant taxi according to the probability and waiting time. Thus they can save their time. The information includes the probability and expected waiting time of taking a vacant taxi on each road nearby. If there are some hotspots where the passenger has a higher probability to take a vacant taxi, we can also give the passenger the information of the hotspots.

Our main contributions of this paper are as follows:

- We propose a locating method to locate on which side of the road segment the passenger is. This is very important because the traffic conditions on two sides of a bidirectional road are different; we need to analyze them separately.
- We propose a model to measure the difficulty of taking a vacant taxi on each road segment by taking the average speed of the taxi, the length of the road segment, and the number of taxis on the road segment into account comprehensively. Based on the model, we predict the probability and waiting time for the potential passenger.
- We detect that if there are some hotspots on the road segment. Then we provide the useful information to the passenger to help him/her do a wise choice.
- We analyze the GPS data of a large number of taxis in Beijing in practical. And then we take the factors of time, date and weather condition into account. Finally we conduct experiments to validate our model, and it performs well.

The rest of this paper is organized as follows. In Section II we give an overview of the related work. Section III formally describes our model which is used to estimate the probability and waiting time for a vacant taxi. Section IV shows how we do our experiments in concretely and the results of our experiments. Finally we make a conclusion and describe the future work.

II. RELATED WORK

In recent years, providing useful service information to passengers to help them take a vacant taxi efficiently has attracted great attention and researchers have done some studies in this area.

Phithakkitnukoon et al. [1] present a predictive model for the number of vacant taxis in a given area. They use the historical data to build the prior probability distributions and employed the method based on the naïve Bayesian classifier. Nonetheless there are a number of limitations of this study that their study is based on 150 taxis in Lisbon, Portugal,
which may not be the true representative of the whole taxi population. In addition, they use their method by dividing the region of the city into 1x1 km² grids that are relatively too large to provide realistic information for the passengers.

Zheng et al. [2] employ the non-homogeneous Poisson process (NHPP) to model the behavior of vacant taxis. They estimate the waiting time by the statistics of the parking time of vacant taxis on the roads and the number of the vacant taxis leaving the roads in history. Based on the estimated waiting time they propose an approach to make recommendations for potential passengers on where to wait for a taxi. However there are still some limitations of this study that the rate function of piecewise linearity in NHPP is too simple for practical situations as the leaving frequency of vacant taxis is continuous. The parameters in their method are regarded as constants during different days; however, these parameters would also be changing slowly as time goes by. When using their method, they don’t consider the effects of the weather.

Qi et al. [3] propose an arriving model to solve the prediction problem of waiting time for a passenger at a given time and spot by mining historical trajectories of taxis. They develop an algorithm to mine taxi traces, simulate the passenger waiting queue for a spot, model the competition of passengers when waiting for a taxi, and build the arrival model of passengers and vacant taxis. By using their model, they can predict the waiting time. However they don’t take the weather condition into account when building their model that affects the accuracy of prediction to some extent.

Yuan et al. [4] present a recommender system to make recommendations for both taxi drivers and people expecting to take a taxi. They establish a probabilistic model to recommend taxi drivers some locations and the routes to these locations, towards which they are more likely to pick up passengers quickly. Similarly they recommend people with some locations where they can easily find vacant taxis. Although their methods could also provide recommendations for passengers, their research mainly focuses on estimating the waiting time to take a vacant taxi, so we must take the length of the road into consideration. Considering the traffic jam in a road segment, the speed of a vacant taxi is also an important factor.

A. TTL Model

Based on the analysis above, we consider a schematic diagram shown in Figure 2. The point S and point E are the starting point and terminal point of road segment $r$ respectively.

We denote the length of $r$ as $L$, and assume the number of vacant taxis in road $r$ at time $t_0$ is $c_0$(the yellow color), the number of taxis entering $r$ is $c_1$ (the red color) from time $t_0$ to time $t_1$, so the time interval $\Delta t$ equals $t_1 - t_0$, namely

$$\Delta t = t_1 - t_0$$  \hfill (1)$$

We denote the number of all the taxis traversing and traversed as $N$, so $N$ equals $c_0 + c_1$, namely

$$N = c_0 + c_1$$  \hfill (2)$$

We define the taxi traffic flow $Q$ during the time interval $\Delta t$ is the number of taxis passing through a road segment during $\Delta t$. So the $Q$ at time $t_0$ is equivalent to the number of taxis newly entering $r$, namely

$$Q = c_1$$  \hfill (3)$$

We denote the average speed of the all the taxis traversing and traversed $r$ during $\Delta t$ as $\bar{V}$, and we assume $Q$
is the same on $r$ (if the road is one-way) or on one side of $r$ (if the road is bidirectional) during $\Delta t$ and the taxis flow is stable during time $\Delta t$. Then we can derive the following formula

$$
\frac{c_0}{c_1} = \frac{L}{\bar{V} \cdot \Delta t}
$$

(4)

Consociating formula (1), (2), (3), (4), we can derive the formula (TTL model)

$$
Q = \frac{N \cdot \bar{V} \cdot \Delta t}{L + \bar{V} \cdot \Delta t}
$$

(5)

Based on the formula (5), we can calculate the taxi traffic flow $Q$, as we can get the value of $N, \bar{V}$ from the historical data of taxis and the $L$ from the road network data. As for $\Delta t$, we can set the value of it according to the real situation.

We all have the experience that the difficulty of taking a vacant taxi varies from morning to night in a day, e.g. there seems to be more vacant taxis in residential areas in the morning but less in the evening. Therefore the time of the day becomes a reasonable indicator for estimating vacant taxis in a given area. Besides the time of day, the day of the week also plays an important role in the distribution of vacant taxis as well. For example, there tends to be more vacant taxis in business areas during the weekdays than in the weekends. Using our TTL model we can calculate a series $Q$ according to the time of the day, the day of the week. We treat the series of $Q$ as a time series, and then we use weighted moving average method (WMA) to predict the current $Q$ after we have calculated the historical taxi traffic flow.

$$
Q_{m+1} = \frac{nQ_m + (n-1)Q_{m-1} + \cdots + 2Q_{m-n+2} + Q_{m-n+1}}{n + (n-1) + \cdots + 2 + 1}
$$

(6)

Here the sequence $Q_m, Q_{m-1}, \cdots, Q_{m-n+2}, Q_{m-n+1}$ represents a series of $Q$ in chronological order that more closely the current time, more larger the weight will be. Namely, the latest time has weight $n$, the second latest $n-1$ and so on.

Considering that the weather condition also plays an important role in measuring the difficulty of taking a vacant taxi. There may be more people going out in a nice sunny day compared to a rainy day. We introduce $\lambda_w$ to weigh the $Q$ when using WMA method, then we derive our improved WMA method below

$$
Q_{m+1} = \frac{nQ_m\lambda_n + (n-1)Q_{m-1}\lambda_{n-1} + \cdots + 2Q_{m-n+2}\lambda_2 + Q_{m-n+1}\lambda_1}{n + (n-1) + \cdots + 2 + 1}
$$

(7)

Here the parameter $\lambda_n$ represents the $\lambda_w$ of the nth day and the value of it can be set empirically or be set according to the following strategy.

In our strategy, we label the weather condition of each day [8] and divide them into four types. As Table 1 shows, we also set different value ($\lambda_{w1}, \lambda_{w2}, \lambda_{w3}, \lambda_{w4}$) to the different type of weather condition. Before using our improved WMA method, we select Table 1 to get the corresponding value of $\lambda_{w1}$ according to the weather condition.

<table>
<thead>
<tr>
<th>Type</th>
<th>Weather condition</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Sunny, Cloudy</td>
<td>$\lambda_{w1}$</td>
</tr>
<tr>
<td>1</td>
<td>Rain, Snow</td>
<td>$\lambda_{w2}$</td>
</tr>
<tr>
<td>2</td>
<td>Hail, Sand storm</td>
<td>$\lambda_{w3}$</td>
</tr>
<tr>
<td>3</td>
<td>Haze, Fog</td>
<td>$\lambda_{w4}$</td>
</tr>
</tbody>
</table>

B. Locating method

Before using our $Q$ model mentioned above, we have to locate the position of the taxi or passenger. We can detect which road segment a geo-position is on by using kd-tree algorithm [9]. As we assume that the two sides of a bidirectional road have different taxi traffic conditions, we have to know the precise information that the passenger or taxi is in $r$.left or $r$.right of the road segment.

We use the Vector cross-product to locate that the passenger or taxi is in $r$.left or $r$.right. The locating model is shown as follows:

![Locating model](image)
As is shown in Figure 3, the point `startNode` and point `endNode` are the starting point and terminal point of a road segment `r` respectively. The point `middleNode` is the middle point of `r`. The point `pointNode` represents that a passenger is waiting a vacant taxi standing on `r`. We denote vector `A` as point `middleNode` to point `peopleNode`, `B` as point `endNode` to point `peopleNode`, we can use `A × B` to judge which side the passenger is standing on.

C. Predicting the probability and waiting time of taking a vacant taxi

After we calculate the taxi traffic flow `Q` on `r` or each side of `r` ( `left` and `right`). We use the variable `Q` to predict the probability and waiting time to take a vacant taxi.

Let `Q_v` be the vacant taxi traffic flow on a road segment `r` during time slot `Δt`. There is no doubt that the bigger the `Q_v` is, the bigger the probability of taking a vacant taxi will be, namely the passenger are more likely to take a vacant taxi.

We define the max taxi traffic flow on all the road segments during `Δt` as `Q_{max}`, and let `P` be the probability of taking a vacant taxi, so we derive the following formula

\[ P = \frac{Q_v}{Q_{max}} \]  

(8)

We can adjust the value of `Q_{max}` according to the actual needs so as to let `P` belong to [0, 1].

The waiting time can also be calculated by the variable `Q`. Let `T_w` be the waiting time that a passenger needs to spend to wait for the first vacant taxi at a given position. There is no doubt that the bigger the `Q` is, the shorter the waiting time will be (In our model we don’t take the waiting queue of passengers into account as we don’t have the relevant data). So it is a negative correlation between the variable `Q` and `T_w`. So we define the formula as follows:

\[ T_w = k \cdot \frac{1}{Q} + \tau \]  

(9)

The variable `k` and `τ` can be trained by the historical data. As the passenger has a tolerance time when waiting for a vacant taxi. We define the tolerance time as `T_{max}` that also represents the upper value of `T_w`. If `T_w > T_{max}`, the infinity will be returned.

D. Detecting and clustering Hotspots

Although the variable `Q` on `r` can be used to measure the difficulty of taking a vacant taxi in a given time, the real condition varies from the starting point to the terminal point of one road. In order to offer the passenger more precise information we detect the hotspot on every road. We define a hotspot as a pick-up point or a drop-off point on a road segment. The hotspots have a higher probability to take a vacant taxi of one road relatively. In this paper we assume that every hotspot is on a road segment and use DBSCAN algorithm [10] to cluster the hotspots on each road.

After a series of works above, given the passenger’s current geo-position, we provide the useful information as flows:

- Locate the road segments nearby his/her current position (within an acceptable walking distance).
- Calculate the current traffic flow `Q`.
- Calculate the probability and waiting time of taking a vacant taxi on each road segment nearby him/her.
- Detect the hotspots on the located road segments.
- Provide all these information to the passenger to help him/her choose a better place to wait for a vacant taxi.

IV. EXPERIMENT

A. Preparation of the GPS data

Our experimental data are the historical GPS trajectory data generated by 12,000 taxis in Beijing in November 2012 [7]. These data cover all the road segments in Beijing. The average sampling interval of trajectory data is 1 minute; the average sampling distance is about 200 meters. The data size is about 50G. Table II shows the fields for each GPS record, along with a few sample entries.

Each GPS point carries the information about the taxi’s identification, location, time, operational status, direction, speed, etc. Let `S = \{s_1, s_2, \cdots\}` represent the set of GPS
points of all the taxis, \(s_k\) represent a GPS point in the format of \((id_{taxi}, t_k, latitude_k, longitude_k, speed_k, direction_k)\).

B. Pretreatment of the road network

The road network data we used in the experiment is the road network of Beijing [6], the road network has over 100,000 network nodes and more than 140,000 road segments that it can cover the entire city.

We extracted the required information from the original road network by using the ArcGIS software [11]. The required information includes the length of road segment, the start point geo-position, and the ending point geo-position. The final format of the road segment \(r\) is like \((id_{road}, lat_s, long_s, lat_e, long_e, length_r)\).

C. Road network matching

We use the pretreated road network to locate the processed GPS data of taxis. We use kd-tree algorithm [9] to locate which road segment \(r\) the taxi traversed and use our locating method to judge which side of the road segment \(r\), the taxi is on \(r.left\) or \(r.right\). Let \(\tau_{dir}\) represents the direction of one road segment, then after road network matching, we can gain the location information in the format of \((t, id_{road}, \tau_{dir}, speed, id_{taxi})\).

D. Model processing

After road network matching, we use our Q model to calculate the variable Q of each road segment. In our experiment, we set the value of time interval \(\Delta t\) to be 10 minutes. Because the data of some roads is very sparse, we only deal with the main roads.

![Figure 4. The prediction value of waiting time](image)

When calculating the traffic flow Q, we use both the WMA method and the improved WMA method to do prediction. The former only distinguishes what day the day is, the latter takes the weather condition into account comparatively. After the prediction of Q, we use the formula (8) and (9) to calculate the probability and waiting time respectively. As shown in Figure 4, we select several road segments at different time in a day, different day in a week to conduct our experiment. The red points represent the real data we test in the field, the blue and green ones represent our prediction whether considering the weather condition. As we can see, the model that takes the weather condition into account performs better than the one that does not. Namely, the former one is closer to the reality.

E. Taxi hotspots detecting and clustering

In this section, we detect the hotspots according to the judgment that if the taxi is becoming occupied status from vacant status (pick-up positions) or becoming vacant status from occupied status (drop-off positions). Then we use the density-based clustering algorithm DBSCAN to cluster the hotspots. In view of that the points have approximately equal waiting time in the same cluster, so calculate the center point of every cluster as the recommended point.

![A) Hotspots distribution](image)

![B) Hotspots clusters aggregated using DBSCAN algorithm](image)

![C) The center points of the hotspots clusters](image)

As it shows in Figure 5, A) shows the hotspots distribution during 9:00am–10:00am nearby Zhichunli. B)
shows the clusters aggregated by using DBSCAN algorithm (different colors represent different clusters). C) shows the center points of the clusters.

When we finish all the above work, we can provide the passenger some valuable information including the probability, waiting time to take a vacant taxi, and the hotspots on the road segment nearby him/her. Figure 6 describes the detailed information can be offered against the scenario mentioned in Section III.

Finally, we conducted in-the-field experience to validate our model. We selected some locations randomly in the main roads of Beijing to do the examination at different time of the day and different day of the week. Our model can approximate the actual value well in a tolerable error of 2 minutes.

![Figure 6. The valuable information](image)

V. CONCLUSION AND FUTURE WORK

To improve passengers’ experience of waiting a vacant taxi is very meaningful by measuring the difficulty of taking a vacant taxi on a road segment at a given time and position. In this paper, we propose a locating method to locate on which side of the road segment the passenger is standing precisely, Then we proposed a model to measure the difficulty of taking a vacant taxi on each road segment by taking the average speed of the taxis, the length of the road segment, the number of vacant taxis on the road segment into account comprehensively. We predict the probability and waiting time based on our model. If there are some hotspots on one road segment, we will detect these hotspots and provide the relevant information of taking a vacant taxi to the passenger so that we can help him/her arrange their itinerary efficiently. We build our model by analyzing the GPS data of a large number of taxis in Beijing. Finally we conducted in-the-field experiments to validate our model, and it performs well.

Our model has a better scalability that is easy to be applied to real-time computing. We will continue to improve our algorithms to optimize the model in order to offer a more accurate service. This model will also be used as a feature of a complete recommendation system (to be done) that will serve as an online service for the potential passenger.

However there are a number of shortages of this work as follows:

We make the prediction on the basis of that only one passenger waits on the road due to the lack of the passengers’ data, but there may be many passengers waiting on the same road segment actually. We are considering take the waiting queue of passengers into account through simulation.

Although we add the factor of weather when improving our model, we assume that the weather condition is changeless throughout the day. Unfortunately, the weather changes with time.

VI. ACKNOWLEDGMENT

This work was based on the First Big Data Innovation and Entrepreneurship Competition[12] organized by China Computer Federation, Chinese Academy of Sciences and the Chongqing Municipal People's Government. This paper was supported by Around Five Top Priorities of "One-Three-Five" Strategic Planning, CNIC(Grant No. CNIC_PY-1403).

VII. REFERENCES