2014 International Conference on Computing, Networking and Communications, Mobile Computing & Vehicle Communications Symposium

An Effective Taxi Recommender System Based on a Spatiotemporal Factor Analysis Model

Yu-Ling Hsueh, Ren-Hung Hwang, and Yu-Ting Chen
Dept. of Computer Science & Information Engineering, National Chung Cheng University, Taiwan
{hsueh@cs.ccu.edu.tw, rhhwang@cs.ccu.edu.tw, wendy10445@gmail.com}

***Abstract*—The taxi fleet management system based on GPS has become an important tool for efficient taxi business. It can be used not only for the sake of fleet management, but also to provide useful information for taxi drivers to earn more profit by mining the historical GPS trajectories. In this paper, we propose a taxi recommender system for next cruising location which could be a value added module of the fleet management system. In the literature, three factors have been considered in different works to provide the similar objective, which are distance between the current location and the recommended location, waiting time for next passengers, and expected fare for the trip. In this paper, in addition to these factors, we consider one more factor based on drivers experience which is the most likely location to pick up passengers given the current passenger drop off location. A location-to-location graph model, referred to as OFF-ON model, is adopted to capture the relation between the passenger get-off location and the next passenger get-on location. We also adopted a ON-OFF model to estimate the expected fare for a trip started at a recommended location. A real world dataset from CRAWDAD is used to evaluated the proposed system. A simulator is developed which simulates cruising behavior of taxies in the dataset and one virtual taxi which cruises based on our recommender system. Our simulation results indicate that although the statistics of historical data may be different from real-time passenger requests, our proposed recommender system is still effective on recommending better profitable cruising location.**

I. Introduction

Due to the dramatic cost down of Global Positioning System (GPS) devices, taxi fleet management system based on GPS has become very popular for taxi companies. By using this system, a taxi company is able to keep track time-stamped GPS trajectories of its taxi cabs. Furthermore, additional information, such as the status of a taxi, include waiting at a stand, cruising, occupied, off shift, can also be tracked. The GPS taxi fleet management can be used not only for the sake of fleet management and security, but also to provide useful information for taxi drivers to earn more profit by mining the historical GPS trajectories and status of taxies. As a consequence, lots of researchers devoted to the research on efficient taxi business, especially the recommender system for taxi drivers under different conditions and objectives [1]–[6]. For a taxi driver, the most concerned topic is likely to be how to maximize his profit. A daily routine of a taxi driver may consist several pairs of cruising time and occupied time. That is, a taxi driver may cruise the road network searching for passengers for a while (which may include waiting at some taxi stands), and then pick up passengers and drive to the designated destination (occupied time). As the passengers get off the taxi, it starts cruising the road network again. It is at this moment that a recommender system could be used to help the taxi driver know where to cruise such that his profit can be increased. The purpose of this work is to recommend a good location for the taxi driver to cruise to such that he can earn more profit than cruise based on his own experience.

Several factors shall be considered for guiding a taxi driver cruising to a more profitable location. First, distance between current location and the recommended location should be short to save time and energy [1], [2]. Secondly, when the taxi arrives the recommended location, the waiting time for next passengers should also be short [2]. Thirdly, if the taxi driver is able to pick up passengers at the recommended location, the fare for the trip is preferred to be large [1]. In the literature, most of the works have considered two of these three factors with different approaches to utilize the historical data. In this paper, we consider all of these factors. In addition, we also consider a fourth factor by mining the relation between the passenger get-off and get-on locations. A location-to-location graph, called OFF-ON model is proposed to capture the relation between the passengers get-off location and the next passenger get-on location. With this model which is calculated based on the historical data, our recommender system is able to know which locations are with high probability to pick up passengers when the taxi driver drop off passengers at a location. Thus, the contribution of this paper is to explore all of these four factors such that the proposed recommender system is effective. In additional, we also analyze, among these four factors, which one are more important than others.

A simulator is developed which simulates cruising behavior of taxies in the data set and one virtual taxi which cruises based on our recommender system. Our simulation results indicate that the revenue earned by the virtual taxi is within the top 1% during the weekdays and top 4% during the weekends when the historical data is used for simulation. The results also show that the revenue earned by the virtual taxi is within the top 6% during the weekdays and top 18% during the weekends when a new data set is used for simulation. We conclude that although the statistics of historical data may be different from real-time passenger requests, our proposed recommender system is still effective on recommending a better profitable cruising location. The remainder of this paper is organized as follows. Section II discusses the related work. The problem is formulated in Section III and our proposed solution is presented in Section IV. Next, Sections V shows the experiment results. Finally, Section VI gives concluding remarks and future directions.

II. Related Work

Most of existing works on cruising location recommender system have only considered two of the four factors aforementioned [1]–[3], [5]. Chang et al. [2] proposed a on-line prediction application to show a context-aware taxi demand system using real maps. They categorized the landmarks or the classes of roads by location ontology, and performed the context-aware pattern mining from taxi request records by adopting different clustering approaches. Through these processes, customer demand can be understood. However, their approach did not validate the demand of the location in different situations in practice. In our work, we find that taxi demand varies in different time slots, and we validated

978-1-4799-2358-8/14/$31.00 ©2014 IEEE

the effect using statistical utilities. Moreira-Matias et al. [3] focused on recommendation for next taxi stand. They adopted time series forecasting techniques to predict the spatiotemporal distribution in real-time. Their results showed that the waiting time to pick-up a passenger could be reduced by 5%. Powell et al. [1] proposed Spatio-Temporal Profitability (STP) map for guiding taxicabs to cruise to more profitable locations. They mainly focused on modeling the profitability of locations. To reduce the time to the recommended location, the STP map only considers a sub-region around the taxi cab's current location which is further divided into a grid of equally sized cells. Time delay between current location and recommended cruising location was also considered. Takayama et al. [5] recommended next waiting/cruising location using paper media which provides taxi drivers with picking-up numbers ranking and/or fee achievement per location only. In [4], Yuan et al. presented a recommender system for both taxi drivers and passengers. For taxi drivers, the system recommends a good parking place with a short waiting time and a long distance for the next trip. Their precision and recall results showed that the method could effectively provide taxi drivers with high revenue positions. However, they did not evaluate by how much they can improve the revenue.

Some other works focused on clustering techniques. For example, Yue et al. [7] explored time-dependent attractive areas and movement patterns, clustering taxi pick-up and drop-off points to discover areas with high taxi demand in five time spans. From the distribution of the clusters in these five time spans, the distributions of crowd flow for a whole day can be observed. Their study suggested that travel distance and destination affect the travel behavior of people. However, their data set is small and only covers the Sunday slots. Zheng et al. [8] regarded places as Hubs and the travel experience of users as Authority, stemming from the web-based model named HITS. A good hub represents a page that points to many other pages, and a good authority represents a page that is linked by many different hubs. Therefore, the ranks of interesting places and influential users are obtained after the simulation processes.

In this paper, our approach is different from the above-mentioned methods in the following aspects. First, we provide a new approach to clustering which is more appropriate for our study. Second, because there are many fluctuating factors regarding the choices of next position when taxi drivers are looking for a new passenger, we provide two new models which take into consideration the importance of these factors, such as the waiting time of the place, the average revenue of the place, the transition probability, and the revenue. Last, we provide a taxi recommender system for taxi drivers and validate the increase in revenue when drivers adopt our system. We also compare our model with the modified HITS model. The average revenue and clusters are regarded as Authority and Hubs, respectively. The rank of locations is provided to the taxi drivers for recommendation in our experiments.

III. Problem Definitions

For taxi drivers, reducing the cruising time on the road and finding the next passenger effectively is the most important task during their operating hours. Taxi drivers tend to wait for passengers in popular positions to reduce their waiting time. Considering the balance of supply and demand, it is not a desirable situation that all taxi drivers go to the same place at the same time. Based on the current time and the current position, our system assists taxi drivers in finding a good potential candidate position by analyzing historical taxi trajectories, thus maximizing their revenue. Many factors affect the choice of the next position after taxi drivers drop off a passenger. These factors that fluctuate with the direction of the taxi drivers include the current time, the distance between the current position and other positions, the probability of passengers waiting at a certain position, and the waiting time of the position. We have designed a model to discuss the relationship of these factors and calculate the scores based on these factors to represent the probability of potential passengers at the positions. Taxi drivers who take the advice of our recommender system can gain higher revenue in a working day. We use the data set from CRAWDAD which contains mobility traces of taxicabs in San Francisco, USA. We have built the recommender system based on analysis of the historical taxi trajectories. There are four dimensions in the data set: latitude, longitude, time stamps, and the status of the taxi cab. Before conducting any further analysis on the subject of location-based services, clustering the GPS coordinates into location is an important step. We select the positions where the passengers get in and out of the vehicles, because we want to understand the movement patterns in the taxi trajectories. After filtering the pick up /drop off positions, we group similar positions together by the process of clustering.

Many clustering algorithms have been developed. The most common clustering approach is K-means [9]. This algorithm aims to divide all items into k groups by the process of iteration until the cluster members no longer change. The parameter k has to be carefully specified. In addition, the obtained result depends on the choice of different initial items in the beginning. Another option is the DBSCAN clustering approach [10], which is a density based clustering algorithm. DBSCAN requires two parameters: epsilon (Eps) and minimum points (MinPts). Unlike K-means, it can deal with arbitrary shaped data sets and outliers better. Both K-means and DBSCAN require some input parameters which affect the output significantly. In the scenario of taxi demand, clusters are used for recommending cruising locations. Thus, the size of clusters shall be controlled to an appropriate value. Hence, we adopt a grid-based clustering approach which groups the pick up/drop off positions into clusters of a fixed size. Specifically, we partition the map into grid cells (or tiles) of a fixed size.

IV. Proposed Solutions

We use historical GPS data to analyze and retrieve the habitation of residents in an area. Temporal and spatial issues are important factors which the taxi drivers may consider in determining passenger distribution. We emphasize on impacts on the movement of the taxi drivers after the passengers are dropped off by analyzing the spatial and temporal factors.

A. Temporal analysis

From our analysis, the daily total revenue on weekdays and weekends are significantly different. Thus, we divide our GPS data into two categories (i.e., weekdays and weekends) so that the properties of taxi demand can be reasonably observed. For instance, requests for taxies are high around office buildings on weekdays. On the other hand, demand for taxies is high around the entertainment centers on weekends. Furthermore, there is great demand for taxies around department stores or business zones during the daytime and around bars at night. As a result, the number of passengers who request taxi services varies with the time and the categories of place. In this paper, we divide a day into 24 time slots of equal length (i.e., the length of a time slot is an hour). The number of divisions affects the performance (i.e., total revenue). In our experiments, we conclude that a higher number of divisions (for example, hourly time slots rather than morning, noon, or afternoon sections) will achieve better performance. Furthermore, we analyze the average waiting time in every place recorded in the historical GPS data. The average waiting time of a place is obtained by dividing the total waiting time by the number of times that passengers get on a taxi in that place. The average waiting time is an important factor that impacts the decision of going to the next pick-up location for taxi drivers.

B. *Spatial analysis*

The first step of the spatial analysis is to conduct the data clustering process on a set of GPS points on the historical GPS trajectories. As aforementioned, we adapt a grid clustering approach and divide the map into fixed square areas. For the positions of GPS trajectories in the San Francisco area, we set 0.2 miles as the diagonal distance for each cluster to ensure that taxi drivers would be in an appropriate cluster for finding potential passengers. After completing the clustering of the GPS data, the distance between two clusters can be obtained. The impact of the distance factor is profound, since fuel costs are directly proportional to the distance driven. We observe that the patterns of the taxi service demand from our historical GPS data can be discovered. For instance, many people go to commercial hotel B after leaving office A. Hence, taxi drivers can go there to wait for the next passenger if they are around office A. The taxi fares which are the source of revenue are charged based on travelling distance. For example, the revenue earned from airport trips is likely to be higher given the fact that an airport is usually located in a suburban area. To sum up the abovementioned spatio-temporal factors, we conclude that there are four main factors that affect total revenue: average waiting time (denoted by T) in each cluster, the distance between two clusters (denoted by D), the average revenue from each cluster (denoted by R), the transition probability which captures the relation of the passenger get-off location and the next passenger get-on location (denoted by P), respectively. We provide two location-to-location (L-L) graph models, On-Off graph and Off-On graph, to obtain more information and to discuss the impacts of these two models further. After observing the routes of taxi drivers, we consider the occupied trips and cruising trips, separately. We build an On-Off graph model as follows. A weighted directed graph G_{on-off} is used to present the moving trajectories of the occupied trips of all taxi drivers. G_{on-off} consists of $(V, E_{on-off}, W_{on-off})$, where V is a set of nodes that represents clusters, E_{on-off} is a set of edges that represents occupied trips, and a weight set W_{on-off} of edges represents the number of transitions between clusters. The edge with a weight of three from point A to B represents that there were three occupied trips from cluster A to B in the historical taxi trajectories. We define the transition probability $p(L_{off}|L_{on})$ as follows. The transition probability represents the probability that passengers get on a taxi at cluster L_i and get off the taxi at cluster L_j. An array P_{on-off} contains each element $p_{ij(on-off)}$, which is given by:

$$p_{ij(on-off)} = \frac{Num(L_{j-off}|L_{i-on})}{Num(L_{i-on})} \tag{1}$$

where $Num(L_{i-on})$ is the number of occupied trips with passengers who get on a taxi at cluster L_i and $Num(L_{j-off}|L_{i-on})$ is the number of occupied trips with passengers who get off a taxi at cluster L_j, after getting into the taxi at cluster L_i. The expected revenue of cluster L_i for a taxi driver is given by:

$$r_i = \sum_{\forall j} p_{ij(on-off) \times r_{ij}} \tag{2}$$

where r_{ij} is the average total revenue from cluster L_i to cluster L_j. r_i is obtained by multiplying the distance from cluster L_i to cluster L_j by the unit fare per mile.

Besides, we consider one more factor based on drivers' experience which is the most likely location to pick up passengers given the current passenger drop off location. We build an Off-On L-L graph model, similar to the On-Off model, as follows. A weighted directed graph G_{off-on} represents the moving trajectories of the cruising trips of all taxi drivers, where $G_{off-on} = (V, E_{off-on}, W_{off-on})$, V is a set of nodes that represents the cluster set E_{off-on}, which is a set of edges that represents cruising trips, and the weight set W_{off-on} of edges represents the number of transitions between clusters. We define the transition probability $p(L_{on}|L_{off})$, which represents the probability of the passengers getting off a taxi at cluster L_i and the taxi driver going to pick up the next passenger at cluster L_j, as follows:

$$P_{ij(off-on)} = \frac{Num(L_{j-on}|L_{i-off})}{Num(L_{i-off})} \tag{3}$$

where $Num(L_{i-off})$ is the number of cruising trips with last passengers who get off a taxi at cluster L_i and $Num(L_{j-on}|L_{i-off})$ is the number of occupied trips with passengers who get on a taxi at cluster L_j given that their previous cruising trips started at cluster L_i.

C. *The training and testing models*

To evaluate the effectiveness of our designed system, we split the GPS data sets into two sets, a training set and a testing set: each data set contains one week of data. First, we built the recommender system using the training set and validated the precision of our system using the testing data. We simulated the revenue of a taxi driver for a whole day using our data set. According to the four factors we have described above, S_k represents the scores of cluster k, calculated by our taxi recommender, which finally finds the best candidate location to seek for the next passenger. The definition of S_k is shown as follows:

$$\begin{aligned} S_k = W_R \times G_{Ri}(k) + W_D \times G_{Di}(k) \\ + W_T \times G_{Ti}(k) + W_P \times G_{Pi}(k) \end{aligned} \tag{4}$$

We vary the parameters of weight to evaluate system performance. The parameters W_R, W_D, W_T, and W_P represent the weight of the average revenue of each cluster, the weight of distance, the weight of waiting time, and the weight of pick-up probability, respectively. If these factor weights are equal to one and the recommender score S_k is close to the sum of all weights, the situation is more suitable for our hypothesis of using cluster k. Before testing the influence of each factor, all factors are normalized based on the gray theory. The variables of $G_{Ri}(k)$, $G_{Di}(k)$, $G_{Ti}(k)$, $G_{Pi}(k)$ will be defined later in this section. Without loss of generality, let us consider a regular sequence in the gray relational space $\{P(X); Q\}$

$$x_i^{(0)}(k) = (x_i^{(0)}(1), ..., x_i^{(0)}(k))$$

If a bigger target value is requested, the following formula is used:

$$x_i^*(k) = \frac{x_i^{(0)}(k) - min.x_i^{(0)}(k)}{max.x_i^{(0)}(k) - min.x_i^{(0)}(k)}$$

Conversely, if a smaller target value is requested, we use the formula:

$$x_i^*(k) = \frac{max.x_i^{(0)}(k) - x_i^{(0)}(k)}{max.x_i^{(0)}(k) - min.x_i^{(0)}(k)}$$

The score $x_i^*(k)$ is the value produced by the gray rational generation, $min.x_i^{(0)}(k)$ and $max.x_i^{(0)}(k)$ are the minimum and maximum in the sequence $x_i^{(0)}(k)$. As a result, the sequences we consider are:

$$G^*_{Ri}(k) = \frac{G_{Ri}(k) - Min.G_{Ri}(k)}{Max.G_{Ri}(k) - Min.G_{Ri}(k)} \tag{5}$$

$$G^*_{Di}(k) = \frac{Max.G_{Di}(k) - G_{Di}(k)}{Max.G_{Di}(k) - Min.G_{Di}(k)} \tag{6}$$

$$G^*_{Ti}(k) = \frac{Max.G_{Ti}(k) - G_{Ti}(k)}{Max.G_{Ti}(k) - Min.G_{Ti}(k)} \tag{7}$$

$$G^*_{Pi}(k) = \frac{G_{Pi}(k) - Min.G_{Pi}(k)}{Max.G_{Pi}(k) - Min.G_{Pi}(k)} \tag{8}$$

$G_{Ri}(k) = (G_{Ri}(1), ..., G_{Ri}(k))$ is a sequence that describes the average revenue from cluster k to all other clusters. $Min.G_{Ri}(k)$ is the maximum revenue and Max.$G_{Ri}(k)$ is the minimum revenue. $G_{Di}(k) = (G_{Di}(1), ..., G_{Di}(k))$ is a sequence that describes the average distance between clusters. $Min.G_{Di}(k)$ is the maximum distance and $Max.G_{Di}(k)$ is the minimum distance. $G_{Ti}(k) = (G_{Ti}(1), ..., G_{Ti}(k))$ is a sequence that describes the average waiting time in the clusters. $Min.G_{Ti}(k)$ is the maximum waiting time and $Max.G_{Ti}(k)$ is the minimum waiting time. $G_{Pi}(k) = (G_{Pi}(1), ..., G_{Pi}(k))$ is a sequence that describes the transition probability of the cluster calculated by our Off-On L-L Graph model. $Min.G_{Pi}(k)$ and $Max.G_{Pi}(k)$ are the minimum and maximum waiting times, respectively. The new sequences $G^*_{Ri}(k), G^*_{Di}(k), G^*_{Ti}(k), G^*_{Pi}(k)$ are between zero to one after normalization. We also aim to find the optimal weight percentage of those factors to achieve an efficient recommender system in the experimental section.

V. Experiments

We used the dataset from CRAWDAD which contains mobility trajectories of taxicabs provided by Yellow Cab in San Francisco, USA, over a period of two weeks. As shown in Table I, given a route distance between the locations where the passengers get on and off a taxicab, we can obtain the fare for the trip.

TABLE I. The charging rules of taxi cabs in San Francisco

San Francisco Taxicab Rates of Fare			
First 1/5 mile	$3.50	Each minute of waiting, or traffic time delay	$0.55
Each additional 1/5th mile	$0.55	Airport Exit Surcharge	$2.00

A. The training model on weekdays

In order to optimize the taxi recommender system, we tested different weight percentages of the factors. As shown in Table II and Figure 1, we observe from the historical data set that the average revenue of 500 taxicabs from Monday to Thursday is $420.82. We simulate the moving paths of the synthetic (virtual) taxi drivers from 20 different starting locations. We conduct the following experiments to test the performance by varying the weights. For example, the setting of the weights (wR=1,wD=1,wT=1, and wP=1) achieved an 89% improvement on the average past revenue, see Figure 1).

TABLE II. Revenues under different weight settings using training data on weekdays

Weight	1,1,1,1	0,1,1,1	0,1,1,2	0,1,2,1	0,1,3,1	0,3,4,1	1,3,4,1
Avg. Revenue	$571.82	$572.43	$505.10	$579.35	$596.28	$587.68	$612.60
Weight	0,1,1,0	0,1,2,0	0,1,3,0	0,2,1,0	0,3,1,0		
Avg. Revenue	$680.42	$606.95	$603.02	$672.16	$648.89		

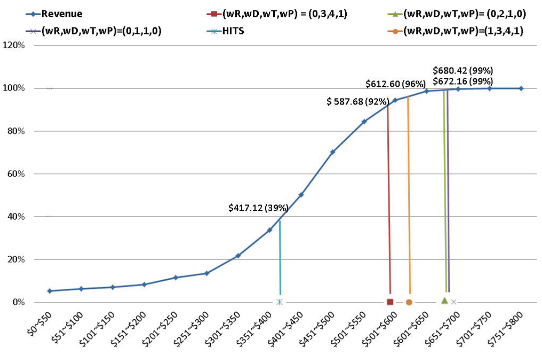

Fig. 1. The cumulative distribution function of revenue using the training data set on weekdays

We adopt the greedy strategy that adjusts only one factor at a time until our system reaches the optimal goal (i.e., maximizing revenue). We then gradually adjust two factors at a time. Finally, we obtain the weight percentage of these factors to achieve the optimal revenue performance. As we can observe from Table II and Figure 1 that the setting of wR=0,wD=1,wT=1, and wP=0 yields the best profit with an average revenue of $680.42. This result indicates that distance and waiting time is more important than the other two factors for the weekday data set. Furthermore, the profit earned by a taxi driver adopting our recommender system is within top 1% of all taxi drivers. Noticeably, the revenue earned by using HITS model is also shown in Figure 1, which recommends taxi drivers to hot spots (short waiting time) without considering other factors. As we can see that the result is not competitive to our approach. In the following, we shall not discuss the results from HITS further as they are not competitive in all of our simulations.

B. The testing model on weekdays

The results of simulation using test data on weekdays are shown in Table III and Figure 2. The average total revenue in the test set is $443.90 in the real situation. When we set the factors according to the best result of the training model, i.e., wR=0, wD=1, wT=1, and wP=0, we can see that the average revenue is $605.93, which is within top 6% of all taxi drivers' earnings. We also investigate other possible settings. In this case, the highest profit ($606.74) can be achieved if we set wR=1, wD=3, wT=4, and wP=1. The results still indicate that distance and waiting time are more important.

TABLE III. Revenues under different weight settings using testing data on weekdays

Weight	1,1,1,1	0,1,1,1	0,1,3,1	0,1,4,1	0,3,4,1	1,3,4,1
Avg. Revenue	469.15	471.01	595.38	546.93	591.86	606.74
Weight	0,1,1,0	0,1,2,0	0,1,3,0	0,2,1,0	0,3,1,0	
Avg. Revenue	605.93	592.72	546.20	590.75	579.69	

C. The training model on weekends

Simulation results using the training data set on weekends are shown in Table IV and Figure 3. The average revenue per taxi is $433.22 in this data set. As we can observe from

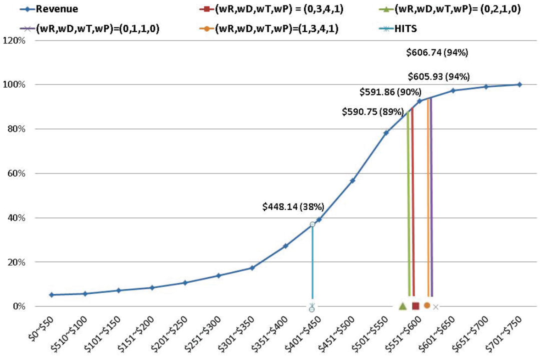

Fig. 2. The cumulative distribution function of revenue using the testing data on weekdays

Figure 3, maximum average revenue, $663.01 (within top 4% of all drivers), can be achieved when weights of four factors are set to wR=0, wD=2, wT=1, and wP =0. An average revenue of $634.5, which is within top 8% of all drivers, can be achieved when weights are set to wR=1, wD=3, wT=4, and wP=1.

TABLE IV. REVENUES UNDER DIFFERENT WEIGHT SETTINGS USING TRAINING DATA ON WEEKENDS

Weight	1,1,1,1	1,3,4,1	0,1,1,1	0,1,1,2	0,3,4,1	0,3,5,1
Avg. Revenue	574.18	634.5	562.99	478.23	619.78	623.13
Weight	0,1,1,0	0,1,2,0	0,1,3,0	0,2,1,0	0,3,1,0	
Avg. Revenue	630.59	614.10	619.75	663.02	633.68	

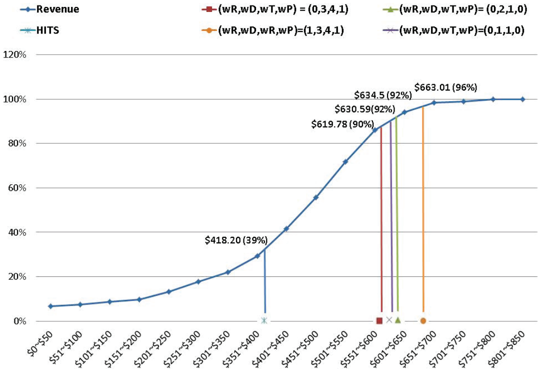

Fig. 3. The cumulative distribution function of revenue using the training data set on weekends

D. The testing model on weekends

Finally, the simulation results using testing data set on weekends are shown in Table V and Figure 4. The average revenue is $491.01 in this data set. The maximum average revenue is achieved, which is $629,92 (within 18% of all drivers), when weights are set to wR=0, wD=1, wT=1, and wP =0. An average revenue of $607.33, which is within top 26% of all drivers, can be achieved when weights are set to wR=1, wD=3, wT=4, and wP=1. Clearly, the results are worse than the other three cases. We conjecture this might due to two reasons. First, this is due to the testing data set is more different from the training data set than the other three cases. This indicates that a more accurate historical data set is important for the recommender system. A way to improve the accuracy is to collect the data for a longer period. Special events of some holidays could also affect the long term prediction. Second, passengers are more crowded on weekends; we can observe from our data set that the probability of picking up passengers is higher, even without the assistance of our taxi recommender system.

TABLE V. REVENUES UNDER DIFFERENT WEIGHT SETTINGS USING TESTING DATA ON WEEKENDS

Weight	1,1,1,1	0,1,1,1	0,3,4,1	0,3,5,1	1,3,4,1
Avg. Revenue	$557.87	$522.26	$619.27	$583.83	$607.33
Weight	0,1,1,0	0,2,1,0	0,1,2,0	0,3,1,0	0,1,3,0
Avg. Revenue	$629.92	$576.10	$562.45	$575.66	$597.20

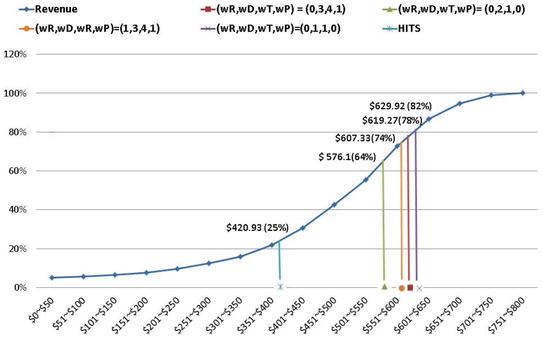

Fig. 4. The cumulative distribution function of revenue using the testing data set on weekends

VI. CONCLUSIONS

In our paper, we investigate four factors on recommending taxi drivers the next cruising location. The four factors are obtained by analyzing the historical data trajectories according to spatio-temporal relation and location-to-location graph models. We evaluated the stability and reliability of our recommender system using real world data sets. Our simulation results show that our recommender system is effective in helping taxi drivers earn more profits even the historical data set may differ from the testing data set. Among these fore factors, we have observed that distance to the next cruising location and waiting time are more important than expected fare of a occupied trip starting at a location and the drivers' experience of picking up passengers from next cruising location based on current location.

REFERENCES

[1] J. W. Powell, Y. Huang, F. Bastani, and M. Ji, "Towards reducing taxicab cruising time using spatio-temporal profitability maps," in *Proceedings of the 12th international conference on Advances in spatial and temporal databases*, Berlin, Heidelberg, 2011, pp. 242–260.

[2] H.-W. Chang, Y. chin Tai, and J. Y. jen Hsu, "Context-aware taxi demand hotspots prediction," *IJBIDM*, vol. 5, no. 1, pp. 3–18, 2010.

[3] L. Moreira-Matias, R. Fernandes, J. Gama, M. Ferreira, J. o. Mendes-Moreira, and L. Damas, "An online recommendation system for the taxi stand choice problem (poster)." in *VNC*, 2012, pp. 173–180.

[4] J. Yuan, Y. Zheng, L. Zhang, X. Xie, and G. Sun, "Where to find my next passenger," in *Proceedings of the 13th international conference on Ubiquitous computing*, New York, NY, USA, 2011, pp. 109–118.

[5] T. Takayama, K. Matsumoto, A. Kumagai, N. Sato, and Y. Murata, "Waiting/cruising location recommendation based on mining on occupied taxi data," *International Journal of Systems Applications, Engineering and Development*, vol. 5, no. 2, pp. 224–236, 2011.

[6] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang, "T-drive: driving directions based on taxi trajectories," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, New York, NY, USA, 2010, pp. 99–108.

[7] Y. Yue, Y. Zhuang, Q. Li, and Q. Mao, "Mining time-dependent attractive areas and movement patterns from taxi trajectory data," in *the 17th International Conference on Geoinformatics*, 2009, pp. 1–6.

[8] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, "Mining interesting locations and travel sequences from gps trajectories," in *the 18th international conference on World wide web*, 2009, pp. 791–800.

[9] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *the fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 1967, pp. 281–297.

[10] M. Ester, H. peter Kriegel, J. S, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.