An Entropy-based Model for Recommendation of Taxis' Cruising Route

Yizhi Liu, Xuesong Wang, Jianxun Liu, Jianjun Wang, Zhuhua Liao and Yijiang Zhao

September 4, 2019

计 算 机 学 报
CHINESE JOURNAL OF COMPUTERS

An Entropy-based Model for Recommendation of Taxis' Cruising Route

Liu Yi-Zhi Wang Xue-Song Liu Jian-Xun Wang Jian-Jun Liao Zhu-Hua Zhao Yi-Jiang

School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan 411201

Key Laboratory of Knowledge Processing and Networked Manufacturing in Hunan Province, Xiangtan 411201

Abstract Recommending an optimal cruising route for a taxi-driver helps he/she save the taxi' idle running time, which can then improve the taxi-drivers' income or reduce the taxi's energy consumption. Mining the optimal knowledge for recommendation from the vast previous drivers' GPS trajectories is a possible way since the trajectories are now easily recorded and kept in databases. Lots of work have been done then. However, existing methods mostly recommend pick-up points for taxis only. Their performance is not good enough since there lacks a good evaluation model for the pick-up points selected. In this paper, we propose a novel evaluation model based on information entropy theory for taxis' cruising route recommendation. Firstly, we select more positional attributes from historical pick-up points in order to obtain accurate spatial-temporal features. Secondly, an integrated evaluation model learning from historical pick-up points is constructed based on the information entropy theory, which is applied to get the future pick-up points. We then design a pruning algorithm to recommend a series of successive points to generate a cruising route for a taxi driver. Experiments are done on a real dataset and the results show that our method can significantly improve the recommendation accuracy of pick-up points, and help taxi-drivers make profitable benefits more than before.

Keywords services recommendation; trajectory data mining; location-based service (LBS); Taxis' cruising route recommendation; information entropy

1 Introduction

Nowadays, the advances in mobile computing and location-acquisition techniques have enabled us a group of location-based services (LBS). LBS based on trajectory data mining mostly aims to alleviate urban traffic congestion and reduce environmental pollution. For instance, trajectory data are applied for road infrastructure monitoring [1][2], traffic status probing [3][4], auxiliary urban planning [5], transportation services improving [6][7][8][9], etc.

Taxis have been considered as a major means of public transportation in modern cities. But taxis spend 35%-60% working time to search passengers by cruising along the roads [10]. Therefore, many researchers focus on pick-up point recommendation from the path-optimization or the profit-maximization perspective. Obviously, there are some defects: (1) they may result in sending all cruising taxis to the same location to compete for the same group of passengers. (2) Single attribute of trajectory data will lead to subjective one-sidedness.

Addressing at above problems, we propose an

This project was funded by the National Natural Science Foundation of China (61872139, 61873316, 41871320)、Provincial and Municipal Joint Fund of Hunan Provincial Natural Science Foundation of China (2018JJ4052)、Hunan Provincial Natural Science Foundation of China (2017JJ2081)、the Key Project of Hunan Provincial Education Department (17A070)、the Project of Hunan Provincial Education Department (17C0646). Liu Yizhi, male, born in 1973, Ph.D., associate professor, is a member of the Computer Society (E200016074S), the main research areas are multimedia content analysis, track mining, location-based services. E-mail: yizhi_liu@sina.cn. Wang Xuesong, female, born in 1995, M.S., no, the main research field is trajectory data mining. E-mail: xuesongwang002@163.com. Liu Jian-Xun (corresponding author), male, born in 1970, Ph.D., professor, Ph.D. supervisor, is a member of the Computer Society (08835D), the main research areas are service computing and cloud computing, E-mail: ljx529@gmail.com

entropy-based model for recommendation of taxis' cruising route. It is able to provide an optimal cruising route recommendation service via measuring the value of historical pick-up points based on information entropy. Our contribution lies in three aspects:

(1) Manifold attributes of taxis’ trajectories are measured by information entropy and are applied to select a sequence of optimal pick-up points for taxis’ cruising route recommendation.

(2) An intercrossing spatio-temporal analysis algorithm is proposed to mine spatial and temporal patterns from historical GPS data.

(3) Our model, the entropy-based cruising route recommendation (ECRR), is evaluated by experiments and outperforms the traditional top-k recommendation method.

The remainder of this paper is organized as follows. Section 2 describes related works. The framework of our approach is put forward in Section 3. Next, we will illustrate our models in Section 4. Section 5 gives our experimental results. And conclusions are drawn in Section 6.

2 Related work

The recommendation of taxi pick-up points can provide drivers with better economic benefits and has received extensive attention from the academic community.

Guojun et al. [11] used a supervised learning model to recommend road clusters to taxi drivers for finding passengers. Artificial neural network is adopted to build a regression model to predict the pickup frequency in a road cluster. Xiangjie et al. [12] through analyzing the quantitative relationship between passengers’ getting on and off taxis, proposed a Time-Location-Relationship combined taxi service recommendation model (TLR). The model utilized Gaussian Process Regression (GPR) method to predict the passenger volume for different regions. Jie et al. [13] propose a data-driven approach to formulate the problem as mining actual reachable region based on real historical trajectory dataset. Yuhong et al. [14] mined the most influential set of locations by traversing the largest number of unique trajectories to find k locations in a given spatial region. In the middle of this process, vertex grouping and best-priority pruning techniques are used to speed up the mining process, and greedy heuristics are used to provide performance guarantees. The paper [15] formulated the data offline processing based on HotSpotScan and Preference Trajectory Scan algorithms and calculated the probability and the waiting time of getting a taxi. Zhang et al. [16] fused the driver's location, distance, and other geographic information into the hidden semantic model to provide the driver with a pick-up area. Hwang et al. [17] analyzes the historical data trajectories according to the spatio-temporal relationship and the location-to-location model to obtain four attributes of the recommendation system for recommending the next cruise location to the taxi driver. Huang et al. [18] used experienced taxi drivers as the learning object in the recommendation system, while considering the prediction of passengers, the prediction of road conditions and the evaluation of benefits.

Yuan et al. [19] proposed a taxi passenger recommendation system based on the pick-up behaviors of high-profit taxi drivers and the mobility patterns of passengers learned from a large number of taxi trajectories. J. Sun et al. [20] developed a novel clustering-based scheme that can exploit multi-source information for taxi pick-up points recommendations. In the literature [21,22], the passenger pick-up point recommendation is obtained by spatiotemporal analysis of the trajectory data. [23] collected useful information from trajectories, then calculated the traffic forces for cruising taxis, based on which taxis are routed to optimal road segments to pick up desired passengers.

Most of the above methods focus on pick-up point recommendation from the perspective of path optimization or profit maximization, which may result in multiple cruising taxis gathering at the same location to compete for the same group of passengers. The singularity and smallness of the pick-up point attributes will also lead to subjective one-sidedness of the recommendation results. Therefore, how to consider the value of the pick-up point as comprehensively as possible is crucial for the taxi pick-up point recommendation.

As a multi-dimensional measurement tool, information entropy has been applied successfully in forecasting and decision-making, system optimization, risk management and analysis, business operation and management [24,25]. Information entropy can be used to determine how much valuable the information is and to measure the weight of each evaluation index in the estimation of information value. It is more meaningful for taxi-drivers to combine multiple indexes to measure the value of pick-up points. Therefore, we propose an entropy-based model for recommendation of taxis' cruising route. By considering multiple evaluation indexes, the pick-up points' selection model is constructed for empty taxis. Information entropy is applied to weight each evaluation index of pick-up points. On this basis, multiple pick-up points are recommended to create the optimal cruising route.

To evaluate the performance of our method, we choose two indexes - the distance after carriage, and the travel time after carriage. They can objectively reflect the value of cruising route recommendation from both spatial and temporal perspectives because they are directly related with the profits of drivers. Comparing with the performance of top-k recommendation, experimental results show that our method can significantly improve the recommendation accuracy of pick-up points and taxi-drivers' profits.

3 The framework of our approach

The framework is shown in Figure 1. Manifold factors are considered to evaluate and further select passenger-finding points via information entropy. The successive points are recommended to taxi-drivers as an optimal cruising route.

Firstly, pick-up points are mined from taxis' trajectories. Pick-up points mean the location carrying passengers. Passengers-finding points mean the gathering areas of passengers which are reckoned as those of pick-up points in this paper. The density of passengers-finding points or pick-up points indirectly reflects the passengers' carrying demands in this region or at this position.

Secondly, we propose a kind of spatial-temporal analysis method based on density clustering. It is able to filter the sparse regions of pick-up points, and capture the passengers-finding points which mean the dense regions of pick-up points. By investigating deeply pick-up points' spatial-temporal distribution characteristics, we can recommend passengers-finding routes for taxi-drivers and support the taxicab dispatching system.

Next, we choose the most determinant attributes by our experiences, and construct an entropy-based model of selecting passengers-finding points. These attributes are consisted of the distance between the taxi's location and the passengers-finding point, the waiting time of passengers-finding points, the carrying probabilities of passengers-finding points, and the expected profits of passengers-finding points. Information entropy is applied to measure their values for taxis.

Finally, we repeat the entropy-based model K-1 time. Each time an optimal passengers-finding point is obtained, and it is used as the current location of taxis for the next round of calculations. Thus, we can get K optimal passengers-finding points by the depth-first method. These points are concatenated as a cruising route recommended to taxi-drivers.

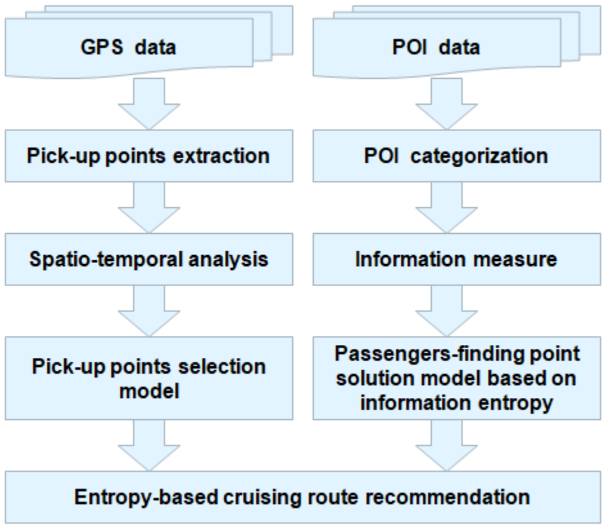


Fig. 1. The framework of our approach.

4 Details of our approach

Information entropy is often used to measure information quantitatively and select uncertainty. When the occurrence probabilities of n different states in the system are denoted as $p_i(i=1,2,\cdots,n)$, the information entropy of the system is:

$$E=-\sum_{i=1}^{n} p_i \ln(p_i). \quad (1)$$

We divide passengers-finding points into two categories: the points with single attribute and those with multiple attributes. The former is only based on the max carrying probabilities or the shortest waiting time, etc. The latter considers the combined effect of two or more attributes.

In this paper, we select four determinant attributes: the carrying probabilities and the waiting time of passengers-finding points, the travel time and the travel distance after carrying. These attributes' effects on finding passengers for empty taxis are uncertain at different location and at different time. So information entropy is used as a quantitative tool to measure uncertainty and trade-off the effects.

4.1 Attributes definition and computation

Taxis' GPS data are firstly preprocessed to distil historical pick-up points and passengers-finding points. The values of the four attributes are computed: the carrying probabilities and the waiting time of passengers-finding points, the travel time and the travel distance after carrying. According to the taxi's current location, the distance and the arrival time to the passengers-finding point can be evaluated. Through the entropy-based model of selecting passengers-finding points, several sequential points are obtained to create a cruising route for recommendation.

The values of the four attributes are computed as follows:

（1）Carrying probabilities of passengers-finding points.

It is denoted as the number of pick-up points in unit time and in unit area. If it is bigger, the chance of finding passengers at this location is bigger.

$$P(i)=\frac{Number(i)}{Area(i)\times T_s}. \quad (2)$$

In the formula (2), $Area(i)$ means the given area range centered on the passengers-finding point, T_s is the given time range. And $Number(i)$ means the number of pick-up points in the spatial and temporal range.

（2）Waiting time of passengers-finding points

It is a time-range for a vacant taxi from waiting to picking up passengers at this location.

$$W(i)=\frac{\sum_{k=1}^{Number(i)} w(k)}{Number(i)\times T_s}. \quad (3)$$

In the formula (3), T_s is the given time range, $Number(i)$ is the number of pick-up points in the spatial and temporal range. And $w(k)$ is the waiting time of the k-th pick-up points in the area $Area(i)$.

（3）Trip time after carrying

In the formula (4), T_s is the given time range, $Number(i)$ is the number of pick-up points in the spatial and temporal range. And $ft(k)$ is the trip time after carrying of the k-th pick-up points in the area $Area(i)$.

$$Ft(i)=\frac{\sum_{k=1}^{Number(i)} ft(k)}{Number(i)\times T_s}. \quad (4)$$

（4）Trip distance after carrying

$$FD(i)=\frac{\sum_{k=1}^{Number(i)} fd(k)}{Number(i)\times T_s}. \quad (5)$$

In the formula (5), T_s is the given time range, $Number(i)$ is the number of pick-up points in the spatial and temporal range. And $fd(k)$ is the trip distance after carrying of the k-th pick-up points in the area $Area(i)$.

Generally speaking, the drivers' profit is proportional with the trip time and the trip distance, and is inversely proportional with the waiting time.

4.2 Pick-up points selection model

As Figure 2 shows, suppose that an empty taxi starts at the point O, the historical passengers-finding points include P_{10}, P_9, P_4, P_1, P_6 five points. We denote it with $X_O=\{P_{10},P_9,P_4,P_1,P_6\}$. The points' location has been figured out by historical data at the current time-slot and within a certain distance threshold.

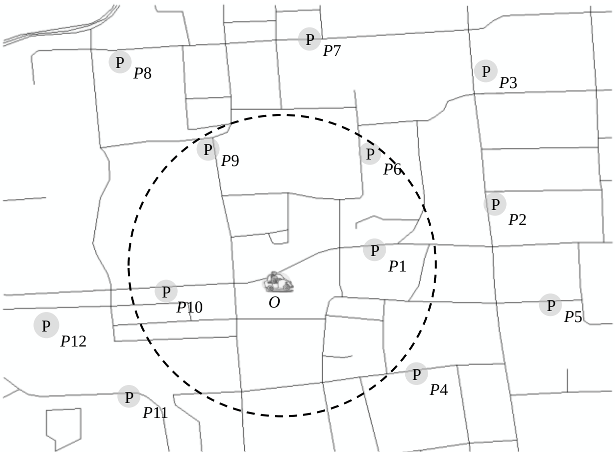


Fig. 2. Pick-up points selection.

When the vacant taxi is located at the point O, there are $k(k \geq 2)$ optional passengers-finding points within the current time-slot. The spatio-temporal features of each passengers-finding point are described by N attribute values. Herein, we choose $N = 5$. The five attributes are respectively carrying probabilities of passengers-finding points, waiting time of passengers-finding points, trip time after carrying, trip distance after carrying, and the distance between the current location and the passengers-finding point. Those passengers-finding points and attribute values are respectively depicted with two sets $X = \{x_1, x_2, \cdots, x_k\}$ and $U = \{p_1, p_2, \cdots, p_N\}$.

The decision matrix A is given as follows. In the matrix, $a_{ij}(1 \leq i \leq k, 1 \leq j \leq n)$ means the j-th attribute value of the i-th passengers-finding point. Different attribute values have different effects on the decisions. To avoid the effects, we employ gravimetric transformation to normalize the decision matrix A, in which r_{ij} are normalized parameters of column vectors according to the equations (7) and (8). For three benefit type attributes, carrying probabilities of passengers-finding points, trip time and trip distance after carrying, we apply the equation (7) to compute r_{ij}. For two cost type attributes, carrying probabilities of passengers-finding points and the distance between the current location and the passengers-finding point, we apply the equation (8) to compute r_{ij}. So the normalized matrix R is created.

$$A = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kN} \end{bmatrix}, \tag{6}$$

$$r_{ij} = \frac{a_{ij}}{\sum_{1 \leq i \leq k} a_{ij}}, \tag{7}$$

$$r_{ij} = \frac{1/a_{ij}}{\sum_{1 \leq i \leq k} 1/a_{ij}}, \tag{8}$$

$$R = (r_{ij})_{k \times N}. \tag{9}$$

4.3 Passengers-finding point solution model based on information entropy

After constructing the selection model of pick-up points, it is necessary to measure the points' values in the decision matrix according to the current location and the current time of taxis. The values are computed in light of the theory of information entropy. The passengers-finding point solution model is given as follows.

Step 1 The entropy E_j is the information entropy of the attribute values p_j in the matrix U.

$$E_j = -\frac{1}{\ln(k)} \sum_{i=1}^{k} r_{ij} \ln(r_{ij}). \tag{10}$$

Step 2 The weight vector $W(w_1, w_2, ..., w_N)$ is computed for the N attribute values.

$$\begin{cases} w_j = \dfrac{(1-E_j)}{\sum_{j=1}^{n}(1-E_j)}, \\ w_1 + w_2 + \cdots + w_N = 1. \end{cases} \tag{11}$$

Where w_j is the weight value of the j-th attribute p_j in U, E_j is the information entropy of p_j in U.

Step 3 The integrated attribute vector $Z(W)$ of $X = \{x_1, x_2, \cdots, x_K\}$ is computed via the following equation.

$$Z(W) = \sum_{i=1}^{k} \sum_{j=1}^{n} r_{ij} w_j. \tag{12}$$

Step 4 To obtain the top-m optimal passengers-finding points, we sort these points by the integrated attribute vector $Z(W)$ and select them from high to low.

$$Z = MAX_m[Sort(Z(W))] \quad (1 \le \mathrm{m} \le \mathrm{k}). \tag{13}$$

The top-m optimal pick-up points can be computed within certain time-slot and in the distance scope via the passengers-finding point solution model based on information entropy. Owing to considering manifold attributes, it effectively avoids the one-sidedness.

4.4 Entropy-based cruising route recommendation

Since the actual situation is dynamic, only a single pick-up point recommended for taxis may lead to the failure of carrying passengers at that position. Therefore, it is very necessary to recommend a suitable cruising route including multiple consecutive pick-up points.

If single factor such as the carrying probabilities is applied for top-k recommendation, taxi-drivers may be puzzled to find the next point when the passenger at the recommended point has been lifted. Therefore, we attempt to recommend not one point but an optimal sequence of pick-up points with the highest probabilities. Figure 3 shows the procedure of recommending the optimal sequence. Passengers release pick-up requirements with the current time and the location at first. Then, the optimal sequence will be constructed via the selection model of pick-up points and the solution model based on information entropy.

According to the current time and the local position, the two models, the selection model of pick-up points and the solution model based on information entropy, are repeated (k-1) times. Each time the previous passengers-finding point is used as the initial point.

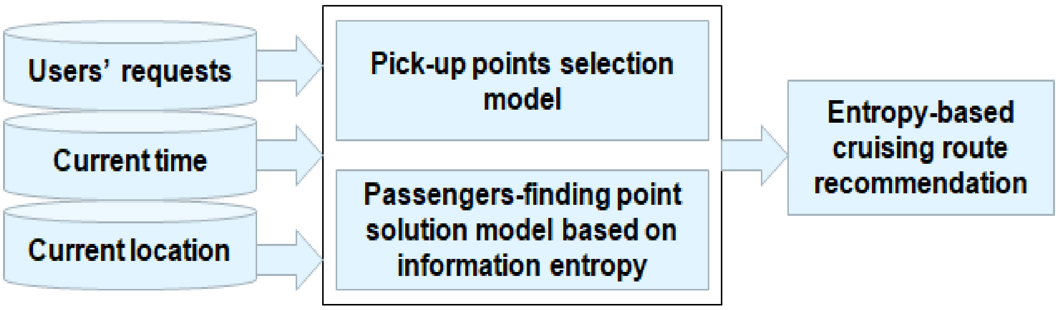


Fig. 3. The procedure of cruising route recommendation.

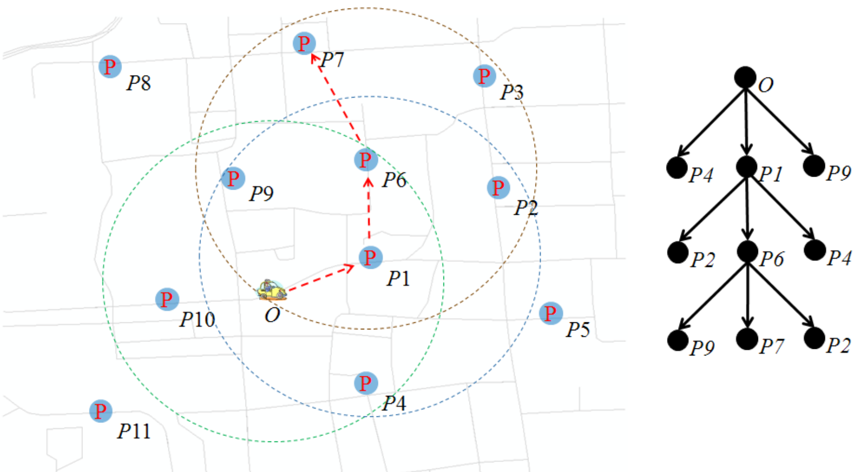


Fig. 4. The procedure of cruising route construction.

Figure 4 shows the procedure of cruising route construction. The location O of a taxi is the initial point. Then the top-m optimal pick-up points $P_{i1}, P_{i2}, \cdots, P_{im}$ can be computed in terms of the time-slot and the distance scope. In turn, one of the previous points is reckoned as the initial point, and then the $m \times m$ optimal pick-up points are computed successively. Thus, a k-layer tree is constructed with optimal pick-up points. Herein, we adopt k=3. Next, the pruning algorithm and the depth-first searching algorithm are explored to select points successively. Consequently, the optimal cruising route is generated and then recommended to taxis.

5 Experiment and analysis

5.1 GPS data preprocessing

The data set used in this paper comes from the open trajectory data set provided by the *GeoLife* project of Microsoft Research Asia, which contains the trajectory data set of 182 users in Beijing from April 2007 to August 2012. The information such as the latitude, longitude and time of the user's location is mainly collected, and the status information is not recorded, that is, the no-load or the state of carrying passengers. Moreover, since the taxi trajectory data is collected by the GPS device, when a special situation occurs, such as signal occlusion, cold start, etc., the signal is weak or no signal may interfere with the data acquisition of the GPS device, Thus there are Abnormal data problem such as causing data loss and data drift. Therefore, it is necessary to preprocess the trajectory and obtain the passenger record point therefrom. The trajectory data preprocessing steps are as follows:

5.1.1 Abnormal point detection

Suppose $T\{p_1, p_2, ..., p_n\}$ is a GPS trajectory sequence, the distance $d(p_{i+1}, p_i)$ between p_i and p_{i+1} should be less than the maximum distance at the maximum speed in unit time. The maximum speed $d_{\max}$ in Beijing is set at 50km/h in this paper.

Therefore, we use the inequation (14) to judge whether the next GPS point is normal or not.

$$0 < d(p_{i+1}, p_i) / (p_{i+1}.t - p_i.t) < d_{\max}. \tag{14}$$

Where $d_{\max}$ is the maximum speed per unit time, $d(p_{i+1}, p_i)$ is the distance between two adjacent GPS points.

5.1.2 Staying area detection

The generation of the staying area includes two situations. On one hand, taxis are parking or waiting for passengers. On the other hand, taxis are cruising. At that time, the moving speed *v* of the moving object at these points is smaller than a threshold θ_v , and continuously moving the time *t* at a speed less than the speed threshold, wherein the time *t* exceeds a given time threshold θ_v, or if the taxi stays in a certain area for more than the time *t*, the area formed by these points is called staying area, and assume that the last point in the staying area is the staying point. The method of reference [26] is used to detect staying area. A staying area $S\{p_n, p_{n+1}, ..., p_i, ..., p_m\} \subset T$ meets:

We reckon the short parking or cruising slowly as waiting for passengers. The thresholds of parking time and cruising velocity are respectively set to be θ_t and θ_v by experiences. In the trajectory sequence $T\{p_1, p_2, ..., p_n\}$, the inequation (15) is applied for stay area detection.

$$\begin{cases} d(p_i, p_{i+1}) / (p_{i+1}.t - p_i.t) < \theta_v, \\ p_m.t - p_n.t > \theta_t. \end{cases} \tag{15}$$

As shown in Figure 5, the trajectory segment $S\{p_3, p_4, p_5, p_6\}$ in the trajectory is a stay area. Only when these points satisfy $v_i < \theta_v, i = 3,4,5,6$ and $p_6.t - p_3.t > \theta_t$.

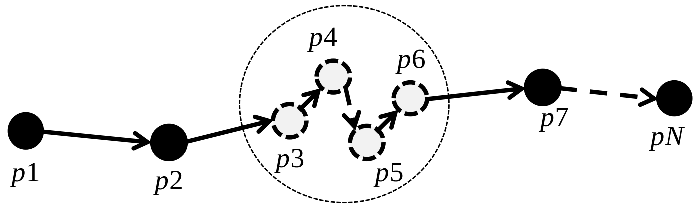


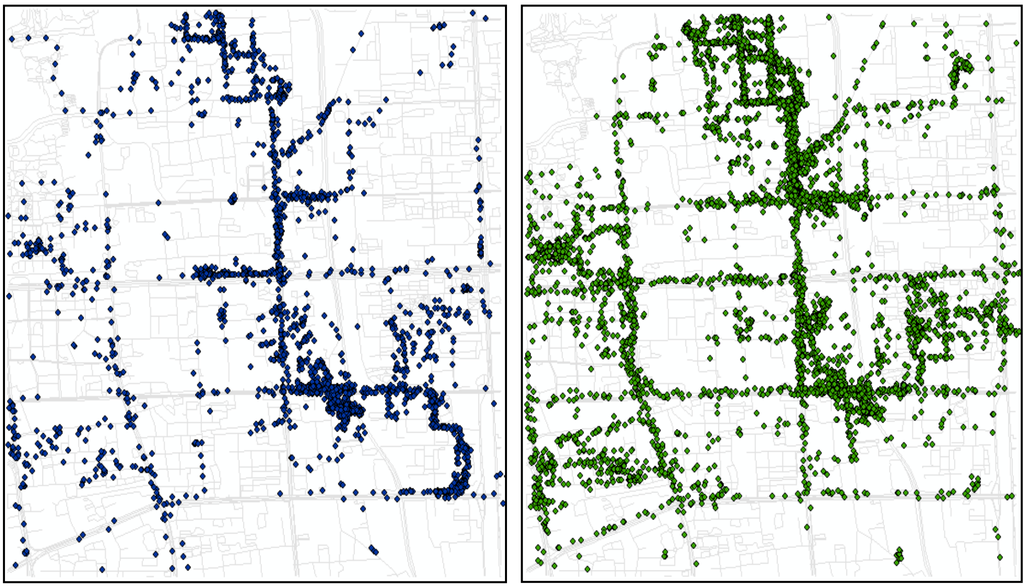
Fig. 5. The procedure of cruising route construction.

5.1.3 Extracting pick-up points

The occurrence of passenger-carrying events usually means that taxis are stationed somewhere or cruise at low speed, so pick-up points can be extracted based on the detection of the residence area. The location where a taxi picks up passengers is usually the last point in the stay area or the end of a low-speed cruise and the starting point of the trajectory. The average speed before carrying passengers is less than the given speed threshold V_θ, and the average speed after carrying passengers exceeds the given speed threshold V_θ. And the driving distance exceeds the given distance threshold δ. The track segment $T\{p_1, p_2, ..., p_i, ..., p_n\}$, p_i is the stop point, and the criterion for defining p_i as the passenger point is:

$$\begin{cases} d(p_0, p_i)/(p_i.t - p_0.t) < V_\theta, \\ d(p_i, p_n)/\ p_n.t - p_i.t) > V_\theta, \\ d(p_i, p_n) > \delta. \end{cases} \quad (16)$$

After extracting pick-up points, you can visualize pick-up points data on the ArcMap, Figure 6 shows the distribution of pick-up points between 8: 00 a.m. and 9: 00 am and 6: 00 p.m. to 7: 00 pm in a district of Beijing. It can be seen from the diagram that the gathering areas of pick-up points are different in different periods, and the degree of aggregation at different locations at the same time is also different. The gathering areas of these pick-up points at different times represent the concentrated positions of passengers in different periods, and the degree of aggregation of passenger points at different locations indicates the difficulty of finding passengers in different locations of no-load taxis.



(a) pick-up points between 8 o'clock and 9 o'clock (b) pick-up points between 18 o'clock and 19 o'clock

Fig. 6. Distribution of pick-up points at different time-slots in the same area.

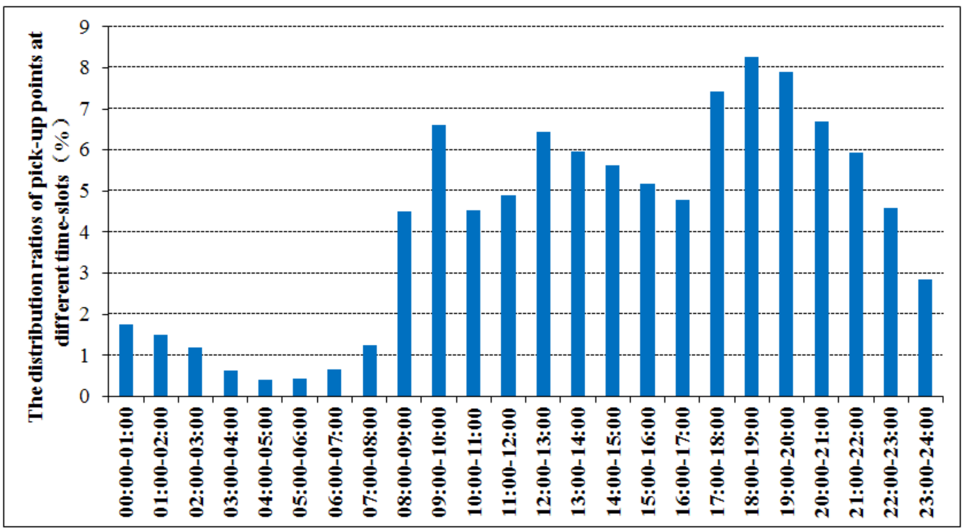


Fig. 7. The distribution ratios of pick-up points at different time-slots.

By dividing 24 hours a day according to the time period, we can find that the distribution of pick-up points in each period has very obvious characteristics. As shown in Figure 7, the distribution ratio of pick-up points in each time of the day can be clearly found that the distribution of passenger points is the lowest between 2 am and 6 am, and the number of passengers is increasing rapidly during each commute time. Passengers-finding points generation and evaluation

5.1.4 Passengers-finding points generation

The gathering area of passengers-finding points is the gathering area of passenger spot, which embodies the characteristics of passenger carrying behavior of taxi and the distribution of people, representing the gathering place of passengers. The pick-up points are the record point in the track, representing the position of the taxi that used to pick up passengers. According to the time attribute and spatial attribute of pick-up points, this paper draws on the cluster analysis method mentioned in [27] and [28], and improves the Spatial Temporal Analysis (STA). In the process of analysis, this paper uses the density-based OPTICS [29] algorithm, mainly because the high-frequency position of pick-up points changes with time, and the spatial distribution of pick-up points in different time periods is different. The algorithm is superior to other algorithms in effect.

The dynamic variability of the high-frequency position of pick-up points is as follows: Different time periods, the location of passengers-finding points is different. During the peak hours of going to work, passengers-finding points appeared in residential areas, apartments and other places. In the peak hours of getting off work, passengers-finding points occurred in office buildings, office buildings and other places. In different urban areas, the occurrence time of passengers-finding points is different. For example, the urban central business district and the science and technology park have different time periods for taxis due to different regional functional properties, so that passengers-finding points of the two regions appear different time periods. .This paper divides the time in hours and divides the day into 24 time slices. The space-time analysis process is as follows.

Procedure 1 Prioritize space and time analysis separately

Procedure 1.1 First Space Last Time (FSLT) Analysis

Priority is given to the spatial distribution characteristics of pick-up points, pick-up points are divided according to the experimental area. Spatial clustering analysis of pick-up points in small areas based on spatial attributes, and then the whole area is analyzed. The results obtained by the spatial attribute analysis are divided according to the time attribute of pick-up points in the Clustering cluster, and the spatial analysis result is retained to the time period when pick-up points ratio of a certain period exceeds a given threshold. A candidate point set $fslt[N]$ can be obtained, each of which is a gathering area of the pick-up points, which represents a gathering area of the passenger to a certain extent.

Procedure 1.2 Analysis of First Time Last Space (FTLS)

This analysis is relatively simple. According to the time aggregation phenomenon of pick-up points, the time analysis is first carried out according to its time attribute, pick-up points are divided into various time segments, and then the spatial clustering analysis is carried out according to the spatial attributes of pick-up points in each time period. A candidate point set $ftls[M]$ can be obtained, each candidate point being a gathering area of pick-up points, which represents a gathering area of the passenger to a certain extent.

First space last time analysis describes the overall spatial aggregation phenomenon of pick-up points, and to some extent compensates for the data sparse problem caused by prioritizing the time attribute. Analysis of first time last space is a better description of the aggregation of the pick-up points on the time attributes. The two processes, time-first analysis and first-time spatial analysis, can be reversed.

Procedure 2 Candidate point filtering

Two candidate data sets are obtained by finite spatial-temporal analysis, which reflect the characteristics of the high-frequency position in pick-up points from different dimensions. In order to obtain the data of the passengers-finding points which represents the passenger gathering place, two data sets should be filtered and the duplicate data should be eliminated. Let the candidate point i come from the candidate point set $fslt[N]$, and the candidate point j from the candidate points set $ftls[M]$, $dist(i,j)$ is the distance between the candidate point i and the candidate point j. When M $dist(i,j)$ is less than a given distance threshold d (such as 50m), it is considered that candidate point i and candidate point j represent the same region and are repeated candidate points.

In this paper, candidate point filtering is performed by voting selection. Taking the distance between pick-up points and the candidate point in the area represented by the candidate point as a voting basis, voting on the above-mentioned repeated candidate points, and then generating a new candidate point according to the distance between the voting result and the candidate point, that is, the final Passengers-finding.

Procedure 2.1 Calculate the score of candidate points i

$$Score(i)=\sum_{1\le k\le K}(1-\frac{dist(i,k)}{dist(i,j)}),dist(i,k)\in[0,d]. \quad (17)$$

Where $Score(i)$ is the voting score of candidate point i, $dist(i,k)$ is the distance between pick-up points k and candidate point i in the area represented by the candidate point, and K is the total number of pass pick-up points in the area represented by the period in which the candidate point is repeated, d denotes a distance threshold for judging whether or not the candidate point is repeated.

Procedure 2.2 Calculate the candidate point j score

$$Score(j)=\sum_{1\le k\le K}(1-\frac{dist(j,k)}{dist(i,j)}),dist(j,k)\in[0,d]. \quad (18)$$

Where $Score(j)$ is the voting score of candidate point j, $dist(j,k)$ is the distance between pick-up points k and candidate point j in the area represented by the candidate point, and K is the total number of pass pick-up points in the area represented by the period in which the candidate point is repeated , d denotes a distance threshold for judging whether or not the candidate point is repeated.

Procedure 2.3 New candidate point generation

After obtaining the voting score of pick-up points to the repeated candidate point, a new candidate point may be generated by using one of the repeated candidate points as a reference position. When the candidate point i is used as the reference position, the new candidate point position is calculated as follows:

$$O=L(i)\sim\frac{Score(i)\times dist(i,j)}{Score(i)+Score(j)}. \quad (19)$$

Where O is the position coordinate of the new candidate point, and $L(i)$ is the position coordinate of the candidate point i.

After obtaining a new candidate point, it is necessary to calculate the relevant area such as the area represented by the new candidate point according to pick-up points of the area represented by the repeated candidate point, and pick-up points in the area is the above $Q[K]$.

5.2.2 Passengers-finding points evaluation

In order to verify that pick-up points obtained by spatial-temporal analysis method can indeed represent the gathering place of passengers. This paper draws on the method of judging the accuracy of docking sites in [27], and uses this method to test the search point passengers-finding points of this paper, based on the fact that taxis usually carry passengers near the point of interest, they believe that docking sites within a 50-meter radius of the point of interest are correct. This paper uses passengers-finding points of each time period as a test point, the point of interest (office building, shopping mall, etc.) on the map and pick-up points are taken as the known points, and the test points are compared with the known points in different periods, so as to judge the correctness of the test points. The spatial-temporal analysis (STA) method of obtaining pick-up points in this paper is compared with the method of obtaining the Carrying-passenger docking point by the hierarchical clustering method in [27].

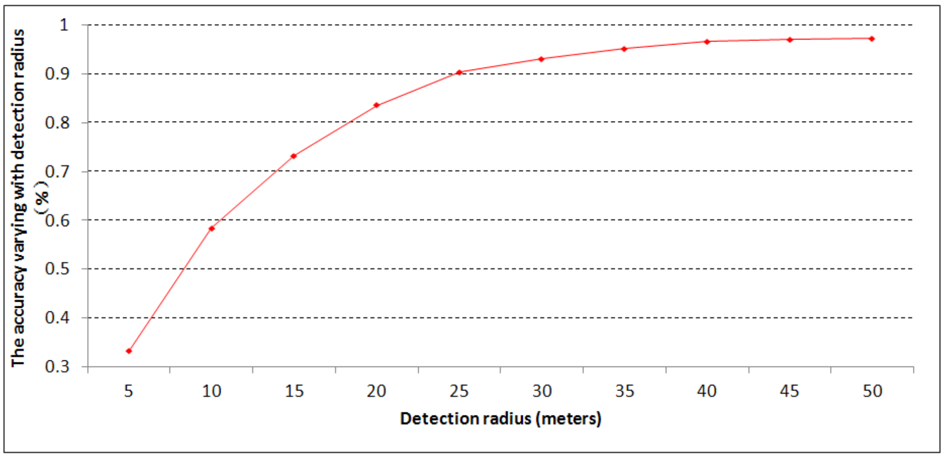


Fig. 8. The accuracy of passengers-finding points recommendation varies with detection radius.

In this paper, the radius of the accuracy of the passengers-finding points is constantly adjusted from 5 meters to 50 meters, and the results are shown in Figure 8. It can be seen that the accuracy of passengers-finding points obtained by the spatial-temporal analysis method reaches 90% at a radius of 25 meters. In the acquisition of pick-up gathering areas or passenger gathering places, the method in this paper is superior to the hierarchical clustering analysis method mentioned in [27]. The spatial and temporal analysis is prioritized separately, which can well reflect the changing characteristics of the high-frequency position of pick-up points, and can more accurately obtain the passengers-finding points representing the gathering place of the passenger.

Table 1. Comparison of spatial-temporal analysis with hierarchical clustering in literature [27]

Method	Radiu of Buffer	Precision
MSRA	50	90.9%
STA	50	98.7%

5.2 Recommendation result evaluation

In order to evaluate the effectiveness of the entropy-based cruising route recommendation (ECRR), the experimental results are compared with the typical Top-K recommendations, and two different values, the ratio of travel time after carrying passengers and driving distance after carrying passengers, are used as the measure of recommended performance.

The typical Top-K recommendation recommends the number of K passengers-finding points with the highest carrying passenger probability near the current location of the taxi. The ratio of driving distance after carrying passengers is the ratio of the driving distance after being carried to the passenger to the total running distance in the current time period. The ratio of travel time after carrying passengers is the ratio of the driving time after being carried to the passenger to the total running time in each period. Figure 9 and Figure 10 respectively compare the ratios of the trip time and the trip distance after carrying passengers between the Top-k approach and our ECRR method.

Figure 9 shows that in the case of more demand for passengers during the day, the carrying-passenger travel time ratio of this method is higher than the Top-K method. In the early morning when the demand for passengers is small, the recommended carrying-passenger travel time ratio of this paper is inferior to Top-K. Because for periods when demand for taxis is small. Because of to the taxi demand small time interval, Top-K gives a high probability of carrying passengers at the location, and relatively more opportunities for carrying passengers. However, this method needs to consider a number of influencing factors, and the road cost largely restricts the choice of these methods for these locations. Therefore, the recommended passengers-finding point is only the nearest optimal, rather than the overall optimal.

Figure 10 shows that the distance of carrying passengers in this method is higher than that of Top-K method in any time period. The reason is that the method takes into account the cost of the journey, while the Top-K method only considers where it is easier to pick up passengers, even the distance between passengers-finding points at which it is recommended to stay is far away. Since the distance after carrying passengers usually indicates the actual payment of passengers, from the perspective of actual revenue, the performance of this method is better than Top-K recommendation, which can bring good benefits to drivers.

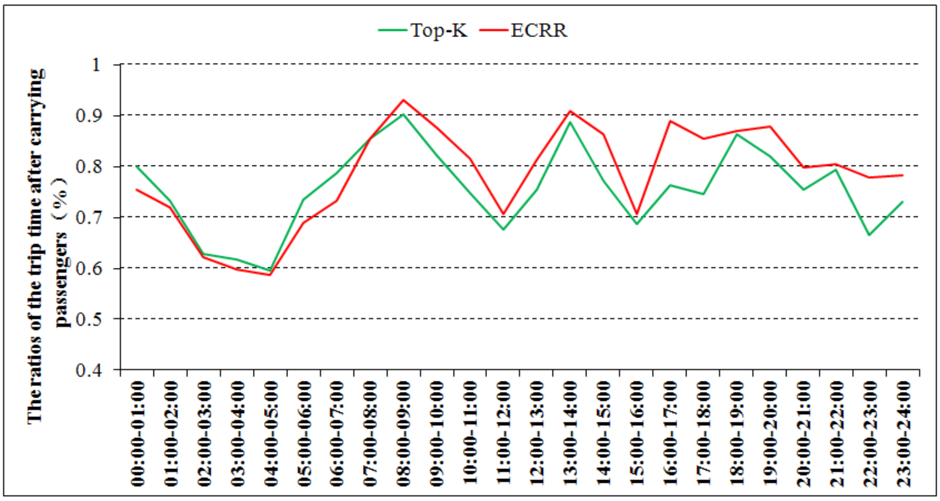


Fig. 9. A comparision on the ratios of the trip time after carrying passengers.

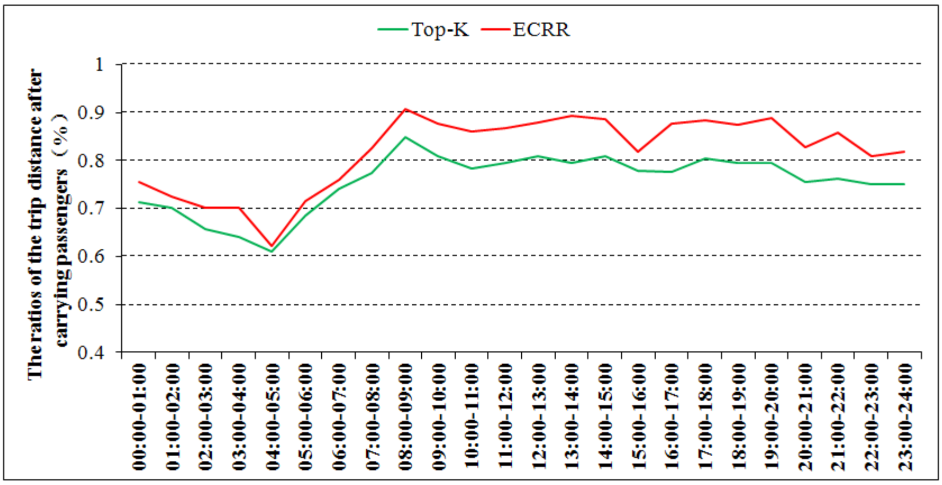


Fig. 10. A comparision on the ratios of the trip distance after carrying passengers.

6 Conclusions

No-load taxis in the process of searching for passengers, different drivers will choose the destination according to different measurement standards, and different choices will lead to different drivers' income differences. How to give a more appropriate passenger seeking route under the consideration of multiple factors, it will undoubtedly increase the income for the no-load taxis' driver. This paper proposes an entropy-based model for recommendation of taxis' cruising route, taking into account the various factors considered by taxi drivers when looking for passengers, and using information entropy to weigh the value of each factor, thereby for the no-load taxi driver to choose the optimal attribute of passenger seeking route. By comparing this method with Top-K recommendation, two different values of driving time and distance after carrying passengers are taken as the evaluation index of recommendation performance. The experimental results show that this method can solve the problem of taxi no-load, save the cost and increase the income of taxi drivers after comprehensive consideration of various factors.

Acknowledgement This project was funded by the National Natural Science Foundation of China

(61872139, 61873316, 41871320), Provincial and Municipal Joint Fund of Hunan Provincial Natural Science Foundation of China (2018JJ4052), Hunan Provincial Natural Science Foundation of China (2017JJ2081), the Key Project of Hunan Provincial Education Department (17A070), the Project of Hunan Provincial Education Department (17C0646).

REFERENCES

[1] Zheng Yu. Urban computing with taxicabs[C]. Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011:89-98.

[2] Cao Lili. And John Krumm. From GPS traces to a routable road map[C], Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2009, 3~12

[3] Chen Yihua. And John Krumm. Probabilistic modeling of traffic lanes from GPS traces[C], Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010, 81~88

[4] Pablo Samuel Castro, Zhang Daqing, Li Shijian. Urban traffic modelling and prediction using large scale taxi GPS traces[M]. Pervasive Computing. Springer Berlin Heidelberg. 2012: 57-72.

[5] Pan Gang, Qi Guande, Wu Zhaohui, et al. Land-use classification using taxi GPS traces[J]. Intelligent Transportation Systems, IEEE Transactions on, 2013, 14(1): 113-123.

[6] Rajesh Krishna Balan, Nguyen Xuan Khoa, Jiang Lingxiao. Real-time trip information service for a large taxi fleet[C]. Proceedings of the 9th international conference on Mobile systems, applications, and services. ACM, 2011: 99-112.

[7] Yuan Jing, Zheng Yu, Xie Xing, et al. Driving with knowledge from the physical world[C]. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2011: 316-324.

[8] Yuan Jing, Zheng Yu, Zhang Chengyang, et al. T-drive: driving directions based on taxi trajectories[C]. Proceedings of the 18th SIGSPATIAL International conference on advances in geographic information systems. ACM, 2010: 99-108.

[9] Qu Meng, Zhu Hengshu, Liu Junming, et al. A cost-effective recommender system for taxi drivers[C]. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014: 45-54.

[10] Dong Hao, Zhang Xuedan, Dong Yuhan, et al. Recommend a profitable cruising route for taxi drivers[C]. Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on. IEEE, 2014: 2003-2008.

[11] Guojun Wu, Yichen Ding, Yanhua Li, Jie Bao, Yu Zheng, Jun Luo. TaxiRec: Recommending Road Clusters to Taxi Drivers Using Ranking-based Extreme Learning Machines [J]. TKDE 2017.

[12] Xiangjie Kong, Feng Xia, Jinzhong Wang, Azizur Rahim, Sajal K. Das.Time-Location-Relationship Combined Service Recommendation Based on Taxi Trajectory Data [J]. IEEE Transactions on Industrial Informatics, 2016.

[13] Jie Bao, Tianfu He, Sijie Ruan, Yanhua Li, Yu Zheng. Mining Spatio-Temporal Reachable Regions over Massive Trajectory Data [C]. In Proceedings of KDD'17, August 13–17, 2017, Halifax, NS, Canada, 11 pages.

[14] Yuhong Li, Jie Bao, Yanhua Li, Yingcai Wu, Zhiguo Gong, Yu Zheng. Mining the Most Influential k-Location Set From Massive Trajectories [J]. TBD 2017.

[15] Xu X, Zhou J, Liu Y, et al. Taxi-RS: Taxi-Hunting Recommendation System Based on Taxi GPS Data[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(4):1-1.

[16] Z. Liao, J. Zhang, Y. Liu, H. Xiao, Y. Zhao and A. Yi, "Fusing Geographic Information Into Latent Factor Model for Pick-Up Region Recommendation," 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shanghai, China, 2019, pp. 330-335.

[17] Hwang R H , Hsueh Y L , Chen Y T . An effective taxi recommender system based on a spatio-temporal factor analysis model[J]. Information Sciences, 2015, 314:28-40.

[18] Huang Z , Shan G , Cheng J , et al. TRec: an efficient recommendation system for hunting passengers with deep neural networks[J]. Neural Computing and Applications, 2018.

[19] Yuan N J , Zheng Y , Zhang L , et al. T-Finder: A Recommender System for Finding Passengers and Vacant Taxis[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10):2390-2403.

[20] J. Sun, C. Guan and M. JinHong, "Exploiting Optimization Mechanism for Pick-up Points Recommendations," 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), Dalian, 2017, pp. 1180-1184.

[21] Zhang M, Liu J, Liu Y, et al. Recommending Pick-up Points for Taxi-drivers Based on Spatio-temporal Clustering[C], International Conference on Cloud & Green Computing. 2012.

[22] Liu Y, Liu J, Liao Z, et al. Recommending a Personalized Sequence of Pick-Up Points[J]. Journal of Computational Science, 2016:278-291.

[23] Lai Y , Lv Z , Li K C , et al. Urban Traffic Coulomb's Law: A New Approach for Taxi Route Recommendation[J]. IEEE Transactions on Intelligent Transportation Systems, PP(99):1-14.

[24] Liu Huawen. Research on feature selection algorithm based on information entropy[D].Jilin University, 2010. (in Chinese)

[25] Xie Hong, Cheng Haozhong, Niu Dongxiao.A Discretization Algorithm for Continuous Sets of Rough Sets Based on Information Entropy[J].Chinese Journal of Computers,2005(09):1570-1574. (in Chinese)

[26] Pedersen Torben Bach. Mining long, sharable patterns in trajectories of moving objects[J]. Geoinformatica. Springer US, 2009, 13(1): 27-

55.

[27] Yuan Jing, Zheng Yu, Zhang Liuhang, et al. Where to find my next passenger[C]. Proceedings of the 13th international conference on Ubiquitous computing. ACM, 2011: 109-118.

[28] Zhang Mingyue. Recommendation for pick-up points and hot zone based on taxi trajectories[D]. Hunan University of Science and Technology, 2013.

[29] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: Ordering Points To Identify the Clustering Structure.[J]. Stanford Research Inst Memo Stanford University, 1999, 28(2):49--60.

Liu Yi-Zhi, born in 1973, Ph.D, associate-professor, His research interests include multimedia content analysis, trajectory mining and location based services.

Wang Xue-Song, born in 1995, M.S. candidate. Her research interests include trajectory data mining and ridesharing system.

Liu Jian-Xun, born in 1970, Ph.D., professor, Ph.D.supervisor. His research interests include service computing and cloud computing.

Wang Jian-Jun, born in 1989, M.S. candidate, the main research direction is track mining.

Liao Zhu-Hua, born in 1977, Ph.D., associate professor, M.S. supervisor. His research interests include Network data acquisition and integration, mobile computing and named data network.

Zhao Yi-Jiang, born in 1977, Ph.D., associate professor, M.S. supervisor. His research interests include VGI and geographic information system.

Background

This paper belongs to intelligent transportation and recommendation system area. Nowadays, the advances in mobile computing and location-acquisition techniques have enabled us a group of location-based services (LBS). LBS based on trajectory data mining mostly aims to alleviate urban traffic congestion and reduce environmental pollution. Recommending an optimal cruising route for a taxi-driver helps he/she save the taxi' idle running time, which can then improve the taxi-drivers' income or reduce the taxi's energy consumption. Mining the optimal knowledge for recommendation from the vast previous drivers' GPS trajectories is a possible way since the trajectories are now easily recorded and kept in databases.

Taxis' Cruising Route recommendation a personalized, location-aware, spatio-temporal recommendation. Therefore, we propose a novel approach of entropy based cruising route recommendation for taxi (ECRR). It is able to provide an optimal cruising route recommendation service via measuring the value of historical pick-up points and the current time and space information based on information entropy.

Firstly, we select more positional attributes from historical pick-up points in order to obtain accurate spatial-temporal features. Secondly, an integrated evaluation model learning from historical pick-up points is constructed based on the information entropy theory, which is applied to get the future pick-up points. We then design a pruning algorithm to recommend a series of successive points to generate a cruising route for a taxi driver. Experiments are done on a real dataset and the results show that our method can significantly improve the recommendation accuracy of pick-up points, and help taxi-drivers make profitable benefits more than before.

The work described in the paper is supported by the National Natural Science Foundation of China (61872139, 61873316, 41871320), Provincial and Municipal Joint Fund of Hunan Provincial Natural Science Foundation of China (2018JJ4052), Hunan Provincial Natural Science Foundation of China (2017JJ2081), the Key Project of Hunan Provincial Education Department (17A070), the Project of Hunan Provincial Education Department (17C0646).