2019 20th IEEE International Conference on Mobile Data Management (MDM)

Passenger searching from Taxi Traces Using HITS-based Inference Model

ZhiFeng Huang
Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
172050025@hdu.edu.cn

Jian Xu
Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
jian.xu@hdu.edu.cn

Guanhua Zhan
Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
172050093@hdu.edu.cn

Ning Zheng
Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
nzheng@hdu.edu.cn

Ming Xu
Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
mxu@hdu.edu.cn

LiMing Tu
Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
tuliming@hdu.edu.cn

***Abstract*—Passenger-searching strategies, as the crowd intelligence of massive taxi drivers, are hidden in their historical GPS traces. Mining traces to understand the efficient passenger-searching strategies can benefit drivers themselves. Traditionally, the research on passenger search strategies from taxi GPS traces is mainly focused on statistical techniques. Although this can improve the ability of drivers to find potential passengers through hotspots recommendation, most of these research still directly use raw GPS data and failed to take drivers' experience into account. Moreover, because driver's experience is behind of raw data, can't obtain directly, so the traditional model is unable to make good use of it during hot spots mining. In this paper, we proposed an inference model based on HITS (Hypertext Induced Topic Search), which perfectly describes the relationship between hot spots and drivers' experience and thus effectively handle the above problem. We first extract hotspots by an innovative P-DBSCAN algorithm based on fuzzy grid partition and match them with corresponding landmarks by a landmark matching algorithm which based on kernel density estimation. Then the HITS-based inference model is used to mine popular hotspots and the most experienced drivers. Finally, we plan an optimal path for drivers. Experimental results demonstrate the efficiency and the ability of this method to provide drivers with better hotspots and hunting sequences recommendation.**

***Index Terms*—Hotspots recommendation, Taxi traces, Taxi trajectory mining, DBSCAN.**

I. Introduction

It has become increasingly common for moving objects (e.g., cars, people) to carry embedded GPS devices, which allow collecting movement data. These GPS trajectories provide us a unique opportunity to understand human behaviors in various situations and exploit the underlying knowledge. However, almost all of these applications still directly use raw GPS data, like coordinates and time stamps, without much understanding of things behind them. so far, they cannot offer much support in giving people valuable information about geospatial locations.

For example, in many cities, taxis are equipped with GPS devices to periodically update their coordinates and passenger status to the central server. These collected GPS trajectories implicitly convey the driver's service strategy, such as where to pick up passengers and how to quickly find the next passenger. Traditionally, research on passenger searching from Taxi GPS Traces has centered around statistical analysis. However, they did not consider the following observation of the relationship between the region and the driver when mining hot spots. 1) The popularity of a hotspot depends on not only the number of passengers picked up by the driver in the area but also these drivers' experiences. 2) The experience of a driver and the popularity of a hotspot are relative values and are region-related. 3) A hotspot' popularity and drivers' experience are interdependent and mutually influence. 4)The popularity of a hotspot is time-dependent. For example, a driver who has served in certain areas like Xiasha, Hangzhou does not know much about the distribution of hotspots in another district of Hangzhou. Or some hot areas such as business centers at night, holidays are more popular than daytime and workdays.

In this paper, based on the daily trajectory of a large number of the taxi, we apply our HITS-based inference model to make full use of the correlation between hotspots and drivers' experiences to find popular hotspots and the optimal hunting sequence. At the same time, we can also dig out more experienced taxi drivers in the region. So this paper strengthens the application of mobile Web model by mining popular hotspots and more experienced drivers in the historical trajectory of a large number of taxi drivers. It provides the possibility of integrating the location-based recommendation service models with mobile Web. The main contributions of this paper are shown as follows:

1) We propose a parallel DBSCAN algorithm based on fuzzy grid partition(FGP-DBSCAN). The algorithm can quickly cluster driver's historical pick-up categories on

2375-0324/19/$31.00 ©2019 IEEE
DOI 10.1109/MDM.2019.00-63

preprocessing data. Each category represents a separate hotspot.

2) We propose a landmark matching algorithm based on kernel density estimation to match a suitable landmark for each hotspot. Then the integration algorithm of k-means and hierarchical clustering is used to build a hierarchical classification tree for landmark data sets. It provides adaptive granularity to represent different regional.
3) It innovatively applies the web page recommendation model HITS to the location recommendation service, which solves the problem that mutually strengthens the relationship between hotspots and experienced drivers. With these knowledge in mind, it plans an optimal hunting path for the driver.
4) We evaluated our algorithms over a large taxi GPS dataset, which was collected by 1600 users over a period of two months in the real world. The number of GPS points exceeded 35 million and its total distance was over 2000,000 kilometers.

II. Related Work

Mining taxi GPS traces has aroused growing attention in many fields, such as ubiquitous computing communities, data mining and intelligent transportation [1], [2]. A large-scale taxi GPS traces helps people deal with all sorts of research issues, such as human mobility understanding [3], [4], [5], [6], [7], traffic prediction [8], [9], city region function identification [10], [11], and taxi/passenger search strategies [4], [5], [12], [13]. In this section, the related work to improve drivers performance in passenger discovery will be briefly introduced.

Most of the relevant papers emphasize identifying and recommending popular pickup areas [4], [7], [12], [13]. Lee et al. [12] made an analysis of the pickup patterns of taxis in Jeju, Korea, and recommended the popular clusters to vacant taxis to reduce the idling time of it. Li et al. [5] predicted the number of pickup events in different hotspots based on the historical information recorded in the taxi GPS traces and has some useful suggestions for vacant taxis drivers was also shared.

Different from the above previous work, we explore the correlation between hotspots and drivers' experience with the help of our HITS-based inference model rather than event-based statistics. Then, we conduct an analysis study to extract the popular and unpopular hotspots at different times and region.

III. Overview of Our System

In this section, we first clarify some terms used in this paper. Then, the architecture of our system is briefly introduced.

A. Preliminary

In this subsection, we will clarify some terms: GPS trajectory, Landmark and Tree-Based Hierarchy.

DEFINITION 1. Taxi Trajectory: *A taxi trajectory Tr is a sequence of GPS points of one trip. Each point p consists of a longitude, latitude and a time stamp p.t, i.e.,* $Tr : p_1 \rightarrow p_2 \rightarrow \ldots \rightarrow p_n, where\ 0 < p_{i+1}.t - p_i.t < \Delta T (1 \leq i < n)$.

DEFINITION 2. Landmark: *Landmark indicates the hotspot of expect the taxi driver to find passengers. A landmark in a narrow sense suggests that there are the most influential building or popular section in the region. Generally speaking, it represents the whole hotspot area.*

DEFINITION 3. Tree-Based Hierarchy TBH: *H is a collection of landmark clusters C with a hierarchy structure* $L.H = (C, L)$, $L = \{l_1, l_2, \ldots, l_n\}$ *denotes the collection of levels of the hierarchy and* $C = \{c_{ij} \mid 1 \leq i \leq |l|, 1 \leq j \leq |C_i|\}$*.means the collection of clusters on different levels. Here,* c_{ij} *represents the jth cluster on level* $l_i \epsilon$ *L, and* C_i *is the collection of clusters on level* l_i.

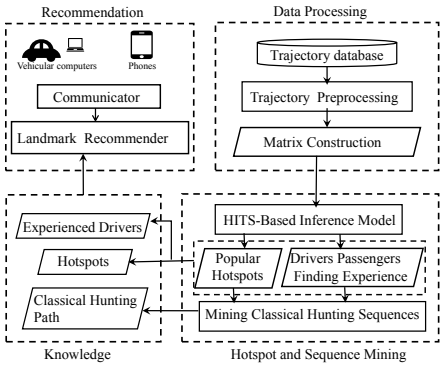

Fig. 1. Architecture of our system

B. Architecture

Fig. 1 shows the architecture of our system, which is comprised of the following three parts: data processing, hotspots and sequence mining, and recommendation. The first two operations can be performed off-line, while the last one should be conducted on-line based on the region specified by a driver.

Data Processing: In the process of data preprocessing, we first clean up the abnormal trajectory, then extract hotspots from the 01 sequences formed by the taxi passenger status, and then match them to the indicative landmarks by a landmark matching algorithm. Finally, according to the personal historical traces sequence of the taxi drivers, a hierarchical diagram of landmarks and a landmark matrix as the HITS-based inference model are established.

Hotspot and sequence: Based on the HITS inference model, we infer the most popular hotspots and the best-experienced drivers in each region. Then we mine the optimal hunting sequence according to the hotness of hotspots and drivers with the best experience. Finally, we plan suitable hunting paths according to the weighted frequent path planning strategy.

Recommendation: To give users better recommendation results, our model combines the experience of drivers and the hotness of hotspots in the area. Given a taxi driver's specific query location, we recommend the most popular hotspots in

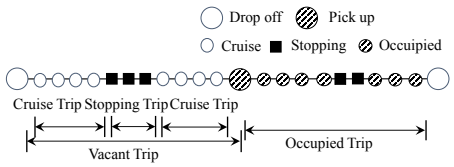

Fig. 2. An example of taxi trajectory

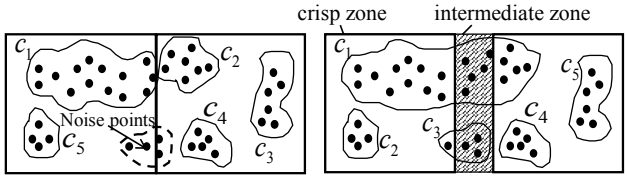

Fig. 3. Example of the sharp boundary problem and our fuzzy grid partition

Algorithm 1: FGP-DBSCAN Algorithm

Input: Dataset: D, Minpts, Eps(ϵ), R
Output: A clustering result: H = $\{h_1, h_2, ..., h_k\}$
1: DB = Data partitioning based on fuzzy grid(D,R,ϵ);
2: C = P-DBSCAN(DB);//The collection of clustering results of each partition
3: **foreach** $c \in C$ **do**
4: **foreach** $cc \in CC$ **do**
5: //CC: A collection of adjacent partition categories for c
6: **foreach** $category1 \in c$-*boundary-category* **do**
7: **foreach** $category2 \in cc$-*boundary-category* **do**
8: **if** *core points of category1* $\bigcap$ *core points of category2* **then**
9: h = $category1 \bigcup category2$;
10: H = H $\bigcup h$;
11: **end if**
12: **end foreach**
13: **end foreach**
14: **end foreach**
15: C = C-CC;
16: **end foreach**
17: **return** H = $\{h_1, h_2, ..., h_k\}$;

IV. Hotspots and Landmarks

In this section, we first describe how to extract hotspots, and then explain in detail how to match a landmark for a hotspot.

A. Extracting Pickup Points

The extraction of a pick-up point depends on the field of "STATE". Based on the field, each unique vehicle's sample points can be assembled as a binary sequence represented by 01...1110...00011. We define "passenger on event" as a shift from 0-1, and "passenger off event" as a shift from 1-0.

In this step, we extract the historical pickup position of the driver in the trajectory to prepare for the hotspot clustering later. The driver's pick-up location, which can be described as the location where "passenger on event" occurs. As shown in the Fig. 2, the position corresponding to the transition state between the empty trajectory and the occupied trajectory is the pick up point we need to extract.

B. Obtaining Hotspots

In this step, we need to cluster more than 40,000 pickup points extracted in the preprocessing process to obtain hotspots. The traditional P-DBSCAN based on clear grid partition to clusters data, although it can accelerate the clustering speed, it often leads to too many clustering categories and introduces new noise points [14]. As shown on the left Fig. 3, This kind of parallel clustering algorithm often divides points belonging to one category into two or more categories because of clear grid partition, and even mistakes some normal points into noise points. As shown on the right Fig. 3, we propose a FGP-DBSCAN algorithm which can solve the hard boundary problem caused by the clear grid partition. As shown in Algorithm 1, the core part of our algorithm is 4 to 17 lines, which mainly describes how we effectively use points in fuzzy space to connect several boundary categories that belong to one category but are separated by data partition. Therefore, this algorithm can greatly reduce the problem of too many classes and noisy points compared with traditional P-DBSCAN.

C. Landmark Matching and Layering

The reason we use "landmark" to model the taxi drivers knowledge is that: The notion of the landmark has strong indicative significance and follows the natural mind-setting of people, and provide drivers a more understandable and memorable presentation of driving directions beyond detailed descriptions. Landmarks are equivalent to hot spots in the following context.

Different from the traditional map matching algorithm, we use the kernel density estimation method to solve the landmark matching problem. In this article, we use candidate landmarks as unknown points of interest, and the set of points in the pick-up category as the set of points of interest that users have visited. The probability of a user accessing a new interest point is the probability that we match the landmark to the hotspot.

Landmark matching model. We use the most popular kernel function $K(x) = \frac{1}{\sqrt{2\Pi}} e^{\frac{-x^2}{2}}$. Given a pick-up category let $L_u = \{l_i\}_{i=1}^m$ be the set of pick up points that have been visited by drivers, m is the total number of pick-up points in this class and D be the set of distances between every pair of the pick up points. For each candidate landmark location that needs to be matched l_j, we define $d_{ij} = |l_i - l_j|$ as the Euclidean distance between l_i and l_j, where $l_i \epsilon L_u$. The kernel density of d_{ij} is defined as:

$$\text{FUNC}(d_{ij}) = \frac{1}{|D|b} \sum_{d' \in D} K(\frac{d_{ij} - d'}{b}). \quad (1)$$

where $b = (\frac{4\hat{\sigma}^5}{3m})^{\frac{1}{5}}$ is an optimized bandwidth. $\hat{\sigma}$ is the standard deviation of the samples in D.

Landmark estimation. Matching Probability of Landmark Location denoted by RATING(l_j) can be calculated by taking the average of all its probability densities, i.e.,

$$\text{RATING}(l_j) = \frac{1}{m} \sum_{i=1}^{m} \text{FUNC}(d_{ij}). \quad (2)$$

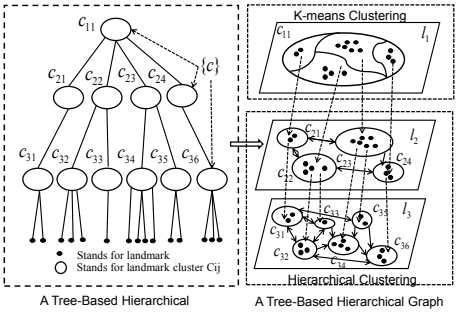

Fig. 4. Building a tree-based hierarchical graph

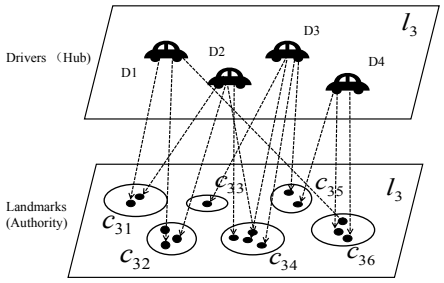

Fig. 5. Our HITS-based inference model

A higher value of RATING(l_j) indicates that the landmark has a higher probability of matching the pick-up category.

Construction of landmark hierarchical graph. As shown in the left half of the Fig. 4, With the hierarchical clustering algorithm, the hierarchical tree of landmark locations in each subregion obtained by k-mean clustering is constructed, and different levels represent different region granularity size. The lower the level, the smaller the granularity. For example, the lowest black dot represents a single landmark(hotspot), while the highest level c is the whole Hangzhou area. Then we are drawing a tree-based hierarchical map combining each taxi's historical trajectory, As shown in the right half of the Fig. 4.

V. Popular Hotspots Inference

In this section, we described how to transform the traditional HITS model which is a search-query-dependent ranking algorithm for Web information retrieval into a HITS-based inference model that fits our scenario.

Based on HITS model, we construct a HITS-based inference model which is suitable for our scenario. We separated two evaluation indexes belonging to a web page and asked them to measure the experience value of drivers and the heat value of hotspots respectively. Using the third level of the picture shown in Fig. 4 as a case illustrates the main idea of our HITS-based inference model. Here, a hotspot(landmark) is a cluster of pick up points. As shown in Fig. 5, similar to HITS, in our model, a hub is a driver who has accessed many hotspots, and authority is a hotspot which has been visited by many drivers. In essence, however, drivers are usually concerned only with the distribution and heat values of the hotspots around their current area and some classic hunting routes within that area. The above hierarchical tree has already helped us to partition data at different regional granularity. In this way, the model can provide drivers with different granularity levels of hot spots and their value distributions.

As shown in Algorithm 2, the matrix M formulated for the case can be represented as follows Equation(3), where all the six clusters pertain to c_{11}.

$$M = \begin{array}{c} \\ D_1 \\ D_2 \\ D_3 \\ D_4 \end{array} \begin{array}{c} \begin{array}{cccccc} c_{31} & c_{32} & c_{33} & c_{34} & c_{35} & c_{36} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 \end{pmatrix} \end{array} \tag{3}$$

In this matrix, an item p_{ij} stands for the times that d_k (driver) has pick-up events to cluster(hotspot) c_{ij}(the jth cluster on the ith level). Such matrixes can be built offline for each non-leaf node based on the drivers' pick-up locations on these hotspots.

Algorithm 2: HITS-based Inference Algorithm

Input: Tree-Based Hierarchy: TBH.
The collection of drivers' pickup points: D.
Output: The collection of drivers' hub scores:H.
The collection of hotspots' authority scores: A.

```
MSet = MatrixBuilding(TBH, D);
foreach M ∈ MSet do
    A = M^T · H;
    H = M · A;
    while EHub> α|| ErrorA > α do
        A_n = M^T · M · A_{n-1};
        H_n = M · M^T · H_{n-1};
        {A_n, H_n} ← Normalization(A_n, H_n);
        {EHub, ErrorA} ← ErrorDetection(A_n, H_n);
    end while
end foreach
return (H, A);
```

Using this powerful iteration inference method, we generate the final scores for each driver and each hotspot. The hotspot with a relatively high authority score is regarded as the more popular hotspot in that category.

VI. Mining the Best Hunting Path

Before planning the optimal hunting path for drivers, we first calculate the scores of each hotspot sequence in a given geospatial area by hotspots' authority and drivers' hub scores. Then, according to the number of times the driver chooses the route and the drivers' experience value, all the routes that can connect the two hot spots are scored. It should be noted that the score of a sequence is the integration of the following three aspects. 1) The sum of the hub scores of drivers who have taken this sequence. 2) The authority scores of the landmark contained in this sequence. 3) These authority scores are weighted based on the probability that drivers would take a specific sequence.

Equation(4) presents the score for any two sequence $L_i \rightarrow L_j$. which includes the following three parts. 1) The authority score of location $L_i(a_{L_j})$ weighted by the probability of drivers' moving out by this sequence $(Out_{L_iL_j})$. 2) The authority score of landmark $L_j(a_{L_j})$ weighted by the probability

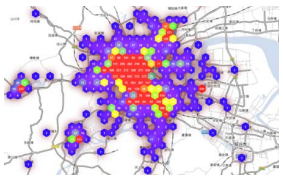

Fig. 6. Weekdays

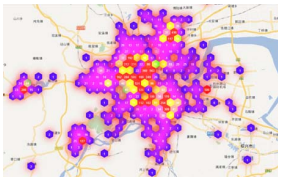

Fig. 7. Weekends

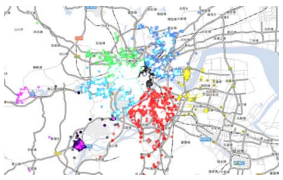

Fig. 8. K-means

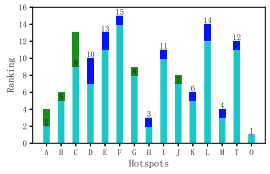

Fig. 9. The popularity ranking of different hotspots in the region

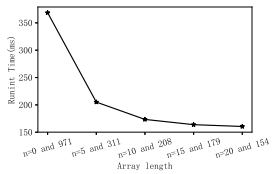

Fig. 10. The influence of the value of n on the running time

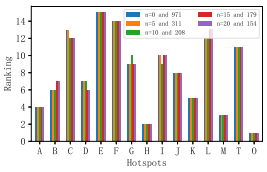

Fig. 11. The influence of the value of n on hot spot ranking results

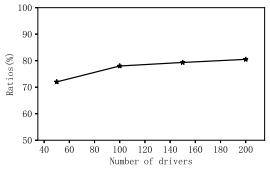

Fig. 12. The most experienced drivers account for the proportion of high-earning drivers

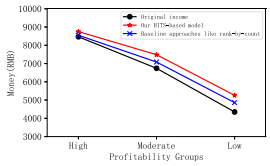

Fig. 13. Driver's Income Gain Change

of drivers' moving in by this sequence $(In_{L_iL_j})$. 3) The hub scores of the drivers $(D_{L_iL_j})$ who have taken this sequence.

$$S_{L_iL_j} = \sum_{d' \in D_{L_iL_j}} (a_{L_i} \cdot Out_{L_iL_j} + a_{L_j} \cdot In_{L_iL_j} + h') \quad (4)$$

Following this method, we can also calculate the score of the sequence $L_i \rightarrow L_k$. Thus, the score of sequence $L_i \rightarrow L_j \rightarrow L_k$ equals to:

$$S_{L_iL_jL_k} = S_{L_iL_j} + S_{L_jL_k} \quad (5)$$

In this paper, we combine the experience value of drivers with the number of times of drivers who choose the route to measure the score of this route. Then recommend the routes with the highest ratings to drivers.

$$S(L_iL_{j\,Pm}) = \sum_{d' \in D'} d'.h \times k \quad (6)$$

Among them, Pm represents the mth path that connects two hotspots, D' Represents the set of drivers who choose the Pm to pass through sequence L_iL_j, $d'.h$ represents the driver's experience value, and k represents the number of times each driver chooses the Pm.

VII. Experiments

In this section, we first briefly describe the experimental data, then report some significant results of HITS-based inference model and finally have some discussion.

A. Data Description

The dataset used in our research consists of temporally-ordered position records collected from about 1644 GPS-enabled taxis within 61 days(October and November 2017), in Hangzhou City, China. The temporal resolution of the dataset is around 40s; thus, theoretically around 2000 GPS points of each car would be recorded in one day(24 h), and the whole volume of the dataset is more than 35 million records. Each GPS point has six attributes, i.e., PLANO, current timestamp(GPS_TIME), current location(longitude, latitude), velocity and car status. The detailed description of the fields is shown in TableI.

TABLE I
Description of the fields of the taxi GPS data

Field	*Value*	*Description*
PLANO	PLANO	9-digit number
LONGI	120.298524	Accurate to 6 decimal places, in degrees
LATI	30.413242	Accurate to 6 decimal places, in degrees
SPEED	29.8	in km/h
STATE	0,1	0, not occupied; 1, occupied
GPS_TIME	20171111000108	14-digit number 11 November 2017 00:01:08

B. The Result of Data Partitioning

In this stage, we systematically analyzed how we divided our data sets. First, we divide the data set into two parts according to the time: weekdays and weekends. Fig. 6 and Fig. 7 show the distribution of pickup events at weekends and weekdays respectively. We can observe that on weekdays, pickup events are mainly concentrated in the downtown and some busy areas. On weekends, pickup events are more discrete than weekdays. The main reason is that people choose to visit scenic spots near their living areas on weekends instead of going to work in the central city. Therefore, it is also a good explanation for the time-dependent nature of the hot spots mentioned above. Then, according to the division of functional areas in Hangzhou, the k-means algorithm is adapted to carry out the spatial division of the data set. Fig. 8 shows the final clustering result of our k-means, and the clustering result is also relatively consistent with the distribution of actual functional areas.

C. Experimental results for our HITS-based inference model

Fig. 9 illustrates the ranking of hotspots popularity. The number marked on each bar is the ranking of the model based on rank-by-count. Among them, the green bar indicates that the hot spot ranking based on our model is backward compared with the rank-by-count method, while the blue bar is the opposite. We can observe that the ranking of hotspots inferred from HITS-based inference model is different from that inferred from the model based on rank-by-count. However, except for very few hotspots, the rankings of other hotspots are only slightly changed, and some even remain unchanged. Although the traditional method can also effectively infer the ranking of hotspots, our inference model can make use of the relationship between drivers and hotspots to give more accurate ranking results.

Fig. 10 and Fig. 11 show the impact of reducing driver array length on running time and hotspots ranking. Among them, n in abscissa indicates the number of passenger pick-up incidents by drivers. By reducing the number of drivers with fewer pick-up incidents, We find that the length of the driver array can be effectively reduced and the running time can be improved. At the same time, it was surprising that there was no obvious change in the ranking of landmarks, only a handful of landmarks swapped rankings with each other when n changes from 5 to 20. This result is probably because the pruned drivers have the least impact on the model speculation because of the sparseness of their pick-up events.

Fig. 12 reveals the proportion of the most experienced drivers we have mined in the highest-income driver group. We can find out that the most experienced drivers accounted for seventy percent of the top 50 high-earning drivers, followed by seventy-eight percent of the top 100 high-earning drivers, and then with the increase of the data volume, the proportion gradually stabilized to eighty percent. The reason for this phenomenon is that the amount of data is small, the proportion is greatly influenced by random factors, and the driver's income depends not only on the passenger's discovery ability but also on many factors, such as path planning and working hours.

In order to prove the efficiency of our model, twenty drivers of different income groups, including high, middle and low-income groups, were randomly selected as the recommended samples. The recommendation experiment shows that it is as shown in Fig. 13. It reveals the changes in income for these three groups after using our model recommendations and model based on rank-by-count. Proposed in this paper models can increase drivers' income, but our model is much better than the other. The incremental income of the low protability drivers' group are higher than those of the high-income group, and then they are higher than those of the middle-income group. The reason for this result should be that the group of high-protability drivers already have rich experience in passenger discovery.

VIII. CONCLUSION

Taxi GPS traces are valuable resources to disclose the crowd intelligence of massive taxi drivers. In this paper, we consider the driver's experience that cannot be obtained from row data when mining hotspots, which is realized through the correlation between hot spots and drivers' experience. Therefore we proposed a HITS-based inference model, through which we can efficiently dig out the region's most popular hotspots and most experienced drivers. Such information can help drivers understand the distribution of hotspots and their hotness in the region and improve their ability to identify potential passengers.

IX. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 61572165).

REFERENCES

[1] P. S. Castro and D. Zhang, "From taxi gps traces to social and community dynamics: A survey," *ACM Comput. Surv.*, vol. 46, no. 2, pp. 17:1–17:34, 2013.

[2] J. Zhang and F. Wang, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[3] H. Cao and N. Mamoulis, "Mining frequent spatio-temporal sequential patterns," in *Proceedings of the 5th IEEE International Conference on Data Mining (ICDM 2005), 27-30 November 2005, Houston, Texas, USA*, 2005, pp. 82–89.

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

[5] X. Li and G. Pan, "Prediction of urban human mobility using large-scale taxi traces and its applications," *Frontiers Comput. Sci. China*, vol. 6, no. 1, pp. 111–121, 2012.

[6] S. Liu and Y. Liu, "Towards mobility-based clustering," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*, 2010, pp. 919–928.

[7] H. A. H. Naji and C. Wu, "Understanding the impact of human mobility patterns on taxi drivers' profitability using clustering techniques: A case study in wuhan, china," *Information*, vol. 8, no. 2, p. 67, 2017.

[8] F. Giannotti and M. Nanni, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *VLDB J.*, vol. 20, no. 5, pp. 695–719, 2011.

[9] Z. Duan and Y. Yang, "Improved deep hybrid networks for urban traffic flow prediction using trajectory data," *IEEE Access*, vol. 6, pp. 31 820–31 827, 2018.

[10] G. Qi and X. Li, "Measuring social functions of city regions from large-scale taxi behaviors," in *Ninth Annual IEEE International Conference on Pervasive Computing and Communications, PerCom 2011, 21-25 March 2011, Seattle, WA, USA, Workshop Proceedings*, 2011, pp. 384–388.

[11] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, 2012, pp. 186–194.

[12] J. Lee and I. Shin, "Analysis of the passenger pick-up pattern for taxi location recommendation," in *NCM 2008, The Fourth International Conference on Networked Computing and Advanced Information Management, Gyeongju, Korea, September 2-4, 2008 - Volume 1*, 2008, pp. 199–204.

[13] Y. Shen and L. ZhaoDBLP:journals/csur/CastroZCLP13, "Analysis and visualization for hot spot based route recommendation using short-dated taxi GPS traces," *Information*, vol. 6, no. 2, pp. 134–151, 2015.

[14] Y. X. Fu and Zhao, "Research on parallel dbscan algorithm design based on mapreduce," in *Advanced Measurement and Test*, vol. 301. Trans Tech Publications, 9 2011, pp. 1133–1138.