This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

A Collaborative Method for Route Discovery Using Taxi Drivers' Experience and Preferences

Zhaocheng He, Kaiying Chen, and Xinyu Chen

Abstract— **This paper presents a collaborative route discovery method that leverages the experience and preferences of taxi drivers in urban areas. The proposed method is mainly comprised of two phases: collaborative preference discovery (CPD) and intelligent driver network generation (IDNG). In the first phase, given an origin–destination (O–D) pair and provided that the cluster is a road segment set within a time-reachable range, we propose CPD which involves cluster-to-cluster retrieval to capture the top-k routes that are not only frequently traversed by taxis but also neighboring to the O–D pair. In the second phase, to support route computation, an IDNG algorithm is devised to generate an experiential graph for each specific O–D pair. In empirical studies, using the period-based experiential route database, sensitivity analysis is employed to select optimal parameters of intelligent driver networks. The results demonstrate that the routes recommended by our collaborative method are much more reliable than those of the shortest-path method with respect to the variance of travel time. Moreover, the recommended routes are traversed more frequently than those of the fastest-path and the shortest-path methods, while the travel time and route lengths of our routes are approximately equal to those of the conventional methods.**

Index Terms— **Intelligent transportation systems, collective intelligence, route planning, taxi trajectory, knowledge acquisition.**

I. Introduction

LOCATION-BASED navigation services play a crucial role in urban travel. With the increasing demand for mobility, the intricate task of devising an effective route is essential for travelers who are asking directions or planning a trip [2], [3]. Typically, taxi drivers, as most experienced ones who have engraved the local network and real-world traffic on their mind, usually use their knowledge to choose routes between an origin-destination (O-D) pair. Recently, researchers have realized that experiential routes collected from taxis can be fruitfully incorporated into route planning. Contrast to the conventional navigation systems computed from static road network, several route learning and computing methods have been developed to foster a manifold of intelligent route planning systems.

In this context, the goal of this study is to provide a route recommendation method using experiential routes which cover taxi drivers' experience and preferences. Several challenges should be tackled, for a given O-D pair, (1) how to measure and formulate the underlying experience and preferences using taxis' historical trajectories; (2) once taxi drivers' experience and preference are decoded, how to structure collective experiential routes into a dynamic O-D purpose-oriented graph; (3) finally, how to calibrate the parameters of the proposed algorithms and further prove that the proposed method can lead to an effective solution for urban route planning problem.

Therefore, in this study, the first task is to develop the Collaborative Preference Discovery (CPD) which is no longer the O-D specified retrieval, but the cluster-to-cluster retrieval where cluster represents a road segment set within a time-reachable range in the road network. To harness the experience and preferences, the collaborative degree calculated by the observations of derived experiential route is critical to select the top-k routes. The second task is to find appropriate O-D purpose-oriented experiential routes in an iterative manner and further generate an Intelligent Driver Network (IDN) which can be treated as the collective intelligence of taxi drivers. In this vein, we propose the Intelligent Driver Network Generation (IDNG) algorithm. Finally, the third task is to select optimal parameters of the proposed method and further compare with the conventional methods (i.e., fastest-path and shortest-path methods). The overall framework of the proposed collaborative method is given by Fig. 1.

The rest of this paper is organized as follows. Section II reviews the related works in this issue. Some basic definitions and data preparation are introduced in Section III. And Section IV describes rationales and details of the proposed route discovery algorithm. In Section V, experiments are reported to present the performance of the compared route planning methods. Further, section VI gives some discussions. Finally, concluding remarks are summarized in Section VII.

II. Related Works

Here, the related works are categorized into three aspects. We first review previous literature on the route recommendation. Then, studies of the latent knowledge and intelligence of taxi drivers are described. Finally, we show that the route learning and computing via historical trajectories can be considered as an available approach to tackle complex route recommendation.

Manuscript received November 17, 2016; revised April 10, 2017 and July 30, 2017; accepted September 2, 2017. This work was supported by the Science and Technology Planning Project of Guangdong province, China, under Grant 2014B010118002. The Associate Editor for this paper was J. Miller. *(Corresponding author: Zhaocheng He.)*

The authors are with the Guangdong Provincial Key Laboratory of Intelligent Transportation Systems, Sun Yat-Sen University, Guangzhou 510006, China, and also with the Research Center of Intelligent Transportation System, Sun Yat-Sen University, Guangzhou 510006, China (e-mail: hezhch@mail.sysu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2017.2753468

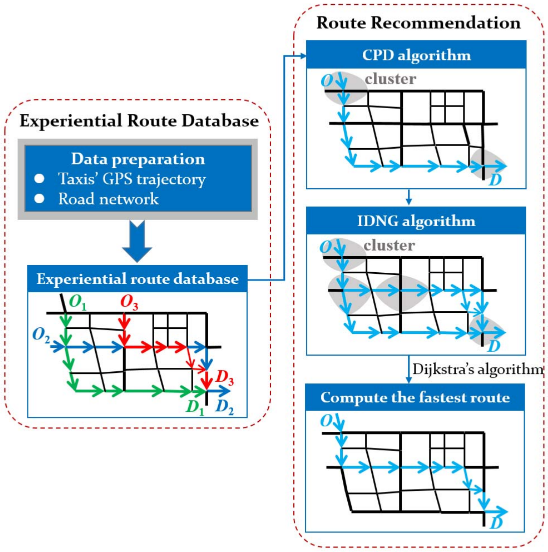

Fig. 1. Overall framework of the collaborative route discovery method.

A. Route Recommendation

With the increasing demand of human mobility, technological innovations such as location-based services hint at a promising approach to explore the power of information-oriented routing [1]. In recent years, a large number of GPS-equipped vehicles such as taxis have become common in urban areas worldwide. In terms of both experience and preference, the intelligence of taxi drivers is hidden in their historical trajectories. Availability of trajectory could foster a number of route recommendation applications which are categorized into follows.

(1) Driving direction enhancement: the fastest route [4]–[6], the shortest route [7], the popular route (i.e., the most frequent route, and sometimes, the top-k frequent routes) [8], [9]–[11], easy-driving route [12], cost-effective route (i.e., shorter travel time, lower costs, and more frequent) [13], [14], and others [15]–[18] are recommended in response to users' (mainly private car drivers [45], [46]) queries.

(2) Taxis' profitable route recommendation [19]–[21]: using drivers' picking-up and dropping-off behaviors, previous studies have attempted to maximize drivers' profit, while some methods intend to recommend locations where passengers can easily find vacant taxis [19].

Relying on crowdsourcing data and crowd knowledge, TripPlanner [24] and CrowdPlanner [25], [26] employed a combination of the location-based social network (i.e., POI network) and taxis' trajectories to achieve personalized, interactive, and traffic-aware routing. Furthermore, [27] aimed to characterize both the dynamics and the uncertainty of road conditions to implement adaptive routing. In [28], [41], and [42], a derivation skyline concept was utilized to retrieve travel routes.

B. Latent Knowledge and Intelligence of Taxi Drivers

Familiarity and Preference: Taxi drivers, as the most skilled and professional drivers in urban areas, are familiar with local road network and traffic. As mentioned in [13], the driving experience of a taxi driver is typically accumulated from thousands of validated route choices in daily life. Based on their familiarity with urban road network, most drivers aim to improve safety and minimizing the possibility of experiencing congestion. Thus, taxi driver will choose a route that offers easy-driving environment [12].

Collective Intelligence: In practice, the latent knowledge and intelligence of taxi drivers are always hidden in large-scale trajectories [7]. Given an O-D pair, [4]–[6] hold that experience can help a taxi driver find the fastest route. Some literature have suggested that, even though the most popular route may not be the fastest, it may offer its own advantages. References [8] and [10] sought out the most popular route, and [9], [11] exploited the top-k popular routes. In this study, we suppose that the collective intelligence discovery from experiential routes can be described by a collaborative manner.

C. Route Learning and Computing

In the route learning, several types of graph extracted from trajectories such as the time-dependent landmark [4], [5], the transfer network [8], the pattern-aware road map [15], the experiential road hierarchy [12], [13], and the grid-based routable graph [9] are constructed to model the collective intelligence of taxi drivers. As real-world traffic and local network characteristics are neglected in the shortest-path method, [13] indicated a new experiential road hierarchy based on travel speed and travel frequency to support routing. Additionally, [9] constructed popular routes from uncertain trajectories, and [15] proposed a novel pattern-aware route discovery framework that considers users' preference.

In the route computing, various algorithms, such as the two-stage routing algorithm [4], [5], the Maximum Probability Product algorithm [8], the constrained breadth-first-search algorithm [15], and the Shortest Path Faster algorithm [7] have been devised to facilitate routing. Although there is no purely objective standard for verification, the routing methods in [4], [5], [12], and [13] were found to suggest routes that were faster than recommended routes of competing methods.

However, to the best of our knowledge, few studies have considered the following two aspects.

1) Collaborative preference. We use the taxis' trajectories with the consideration of its travel frequency in specific time period and spatial distance to a certain O-D pair. To some extents, O-D purpose-oriented and cluster-to-cluster route discovery can covert collective intelligence into a collaborative formula while capturing top-k routes.

2) O-D purpose-oriented partial graph. Once experiential routes are well-organized, we can achieve appropriate route recommendations by only considering refined partial graph with highly collaborative routes rather than a global graph.

III. Problem Definitions

A. Experiential Route Database (ERDB)

Define $\bar{t}_l$ as the average travel time of road segment l and f_l as the frequency (i.e., total number of taxi trips traversed l), the travel time of an experiential route is the addition of travel

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

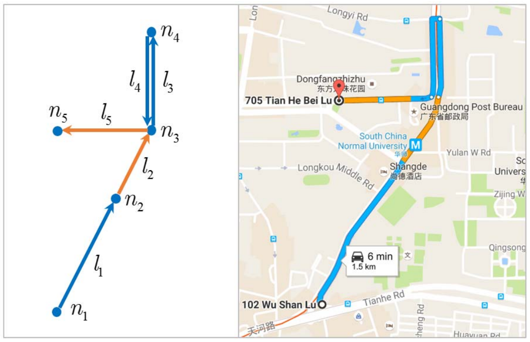

Fig. 2. Representative example of an experiential route.

times of their road segments, i.e.,

$$\bar{t}_{route} = \sum_{l \in route} \bar{t}_l \tag{1}$$

where $l \in route$ represents road segment l in the experiential route. $\bar{t}_l = \frac{1}{f_l} \sum_{i=1}^{f_l} t_l$ is the average travel time of road segment l. The frequency of experiential route (i.e., the total taxi trips) is denoted by f_{route}.

B. Experiential Route Database (ERDB)

Time-evolving traffic patterns may cause a taxi driver make different route choices. Thus, we construct a period-based database. In this study, days of the week are roughly divided into weekdays and weekends, further, time periods of the day include morning peak hours, evening peak hours and other off-peak hours.

In the ERDB, raw GPS trajectory data has converted into a sequence of road segments. If the route frequency of a route like $\{l_1, l_2, l_3, l_4, l_5\}$ as exemplified by Fig. 2 is $f_{route} = 248$, this means that there are 248 trips along this route.

C. Collaborative Preference Discovery (CPD)

When using taxis' trajectory, an inevitable issue is sparseness, meaning that it is impossible to always have experiential routes matching to a given O-D pair directly. However, by means of the collaborative preference discovery, routes between two physical regions can be retrieved. For example, in Fig. 3(a), three frequent routes are illustrated, including r_1 (from l_1 to l_4), r_2(from l_2 to l_4), and r_3 (from l_2 to l_3). If there is no complete experiential route r_1 for the O-D pair (i.e., l_1-l_4), then with the cluster-to-cluster retrieval, neighboring routes (i.e., r_2 and r_3) are potential options.

The collaborative preference is lied in experiential routes with similar travel purposes. Thus, we iteratively traverse road segments in previously derived experiential routes to capture more experiential routes with collaborative preference. Taking Fig. 3(b) for example, the road segment l_6 in r_1 forms c_{e1} which advise a candidate $\{l_1, l_5, l_6, l_7, r_4, l_3, l_4\}$ from l_1 to l_4.

After briefly introduced the main ideas of our collaborative method. Several terms and definitions are detailed as follows.

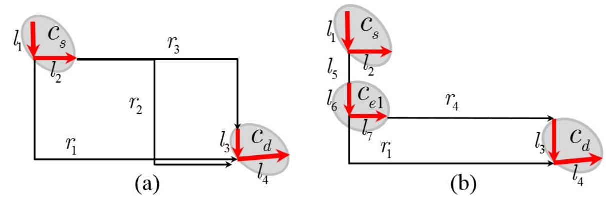

Fig. 3. Application of collaborative preference. (a) Cluster-to-cluster retrieval. (b) Iterative collaborative preference discovery.

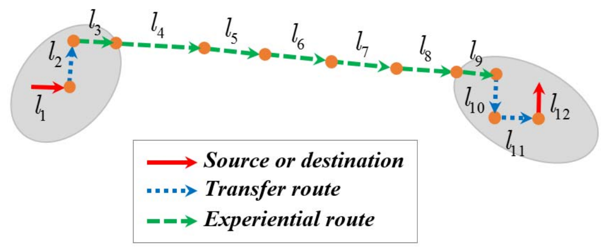

Fig. 4. An example of an experiential route and their transfer routes.

1) Cluster: A cluster can be regarded as the road segment set within a time-reachable region. Specifically, as shown in Fig. 3, the clusters are classified into three types, i.e., source cluster (c_s), destination cluster (c_d), and extended cluster (c_e). Starting from a source, the road segments in the source cluster and the extended cluster are reachable within a prescribed time threshold. Likewise, the destination is required to be reachable within the same time threshold starting from any road segment in the destination cluster.

2) Time-Reachable Threshold t_c: To avoid an abnormal cluster arising via the inclusion of too many road segments, we set a time-reachable threshold t_c as a condition. In the following studies, all clusters are restricted by the same t_c.

3) Cluster-to-Cluster Retrieval: In Fig. 3, the cluster-to-cluster retrieval from c_s to c_d covers r_1, r_2, and r_3, while r_4 is the retrieval result from c_{e1} to c_d. Given a route ranking scheme, the top-k routes can be selected from the source cluster c_s (or extended cluster c_e) to the destination cluster c_d.

4) Query Experiential Route R_q: A query experiential route is one that proceeds from c_s to c_d such as r_1, r_2, and r_3 shown in Fig. 3.

5) Extended Experiential Route : R_e: The extended experiential route is captured from an extended cluster to the destination cluster. In Fig. 3(b), r_4 is an extended experiential route from c_{e1} to c_d.

6) Transfer Route: R_t: From c_s(or c_e) to c_d, the start road segment (or end road segment) of an derived experiential route may not be the source (or the destination). In this vein, a transfer route is necessary to connect the source (or destination) with the start road segment (or end road segment), and the experiential route can be completely matched to a journey. Here, the transfer route is the fastest option and has been previously traversed by taxis. In Fig. 4, l_2 is the transfer route for the experiential route in source cluster, and $\{l_{10}, l_{11}\}$ is the transfer route in destination cluster.

7) Transfer Time: Time interval from the source to the start road segment of an experiential route in the source cluster is denoted by t_s, and from the end road segment of an experiential route (i.e., query experiential route or extended experiential route) to the destination in the destination cluster is denoted by t_d. Similarly, we define the transfer time t_e for extended cluster.

8) Collaborative Degree C_{route}: To strike a balance between the transfer times t_s, t_d and frequency f_{route}, we define a ranking score for experiential route, collaborative degree (similar to TF-IDF [33]), which is formulated as Equation (2).

$$C_{route} = f_{route} \cdot \log\left(\frac{2\max\{t_c\}+1}{t_s+t_d+1}\right) \tag{2}$$

where $\max\{t_c\}$ is the maximum of prescribed t_c, and the denominator t_s+t_d+1 is used to avoid zero appearing in the logarithmic function. When calculate the collaborative degree of an extended experiential route, t_s can be replaced by t_e. If the percentage $p = \frac{t_s+t_d+1}{2\max\{t_c\}+1}$ is regarded as an experience measure, then $\log(1/p)$ can be interpreted as experiential information. Shorter transfer times t_s and t_d, along with higher travel frequency f_{route}, tend to contribute more experiential information of the experiential route.

9) Top-k Routes: The top-k routes refer to the routes with top-k largest collaborative degree values among total experiential routes from c_s(or c_e) to c_d.

10) Intelligent Driver Network (IDN): An IDN is a directed graph $G = (V, E)$ where $V$and E denote intersection set and road segment set, respectively. Note that three route types, including query experiential routes, extended experiential routes, and their corresponding transfer routes, are leveraged to construct an IDN. Hence, the IDN can be treated as an experiential road network.

D. Sensitivity, Efficiency, Reliability and Effectiveness

The performance of route recommendation in the IDN is sensitive to the network structure. We attempt to find a well-structured IDN with the optimal time-reachable threshold t_c and rank-threshold k.

Computing the fastest route, the computational complexity of Dijkstra's algorithm is proportional to the number of road segments or intersections in a road network. Thus, the efficiency can be measured by connection scale, counting the connections for all traversed road segments in the IDN. For instance, in Fig. 2, the connection scale is 4, essentially, $\{l.1, l.2\}$, $\{l.2, l.3\}$, $\{l.3, l.4\}$, and $\{l.4, l.5\}$ are the connections.

The reliability is quantified by the variance of travel time of the recommended route (i.e., Equation (4)). Obviously, the higher the variance is, the less reliable the route will be.

$$Var(l) = \frac{1}{f_l - 1}\sum_{i=1}^{f_l}\left(t_l^i - \bar{t}_l\right)^2 \tag{3}$$

where t_l^i is the travel time of road segment l in ith trip. To avoid confusion, in Equation (4), the notation Var_{route} represents the average over $Var(l)$.

$$Var_{route} = \frac{1}{q}\sum_{l\in route} Var(l) \tag{4}$$

where q is the number of road segments for the recommended route.

The effectiveness is measured by frequency, travel time, and route length. The average frequency of the recommended route is given by Equation (5).

$$\langle f_{route}\rangle = \frac{1}{q}\sum_{l\in route} f_l. \tag{5}$$

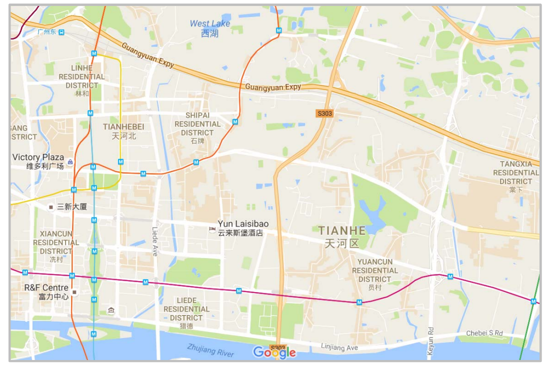

Fig. 5. The Tianhe Central Commercial District in Google Maps.

TABLE I

DATA PREPARATION AND EXPERIENTIAL ROUTE DATABASE

Pattern	Raw Sample	Raw Average Frequency	Updated Sample	Updated Average Frequency
WDMP	79,619	2.02	1,589,054	7.66
WDOP	395,876	3.16	7,043,065	13.48
WDEP	75,216	1.87	1,732,509	6.27
WEMP	23,960	1.54	418,965	6.57
WEOP	153,280	2.42	2,291,424	10.48
WEEP	34,184	1.59	635,281	5.58

[a]WDMP: weekday morning peak hours; WDOP: weekday off-peak hours; WDEP: weekday evening peak hours; WEMP: weekend morning peak hours; WEOP: weekend off-peak hours; WEEP: weekend evening peak hours.

IV. ROUTE LEARNING AND COMPUTING

In this section, we will describe the CPD and IDNG algorithms in detail, and present the process of route learning and computing meanwhile.

A. Data Preparation and Database Construction

The real trajectory data was collected over a period of around 4 weeks from Tianhe Central Commercial District (i.e., an area of approximately 7km×5km=35km^2 shown in Fig. 5) in the city of Guangzhou, China. The number of road segment and intersection are 410 and 275, respectively. Table I shows the patterns and raw numbers of the trajectory samples, and their average frequency that are used for database construction. It is worth noting that there are more than 20, 000 taxis in Guangzhou city, enough taxi drivers and trajectory data in this area are available to support this route discovery experiment.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

In addition, other considerations for choosing this area are given as follows. First, the complex traffic conditions and frequent congestion in the urban central areas cause the harness of experience of route decision from experienced drivers to be valuable. Second, taxi drivers tend to serve passengers in their profitable central areas, as a result, more trajectory data can be observed in central areas than other areas to support the route discovery task.

For the reason that knowledge discovery is significantly dependent on the sample of taxis' trajectory, thus, regarding the fact that an experiential route may include some underlying sub-routes, we extract the sub-routes from only parts of the route that traversed the study area . For instance, given a route $\{l_1, l_2, l_3, l_4, l_5, l_6, l_7\}$, suppose that the road segments l_1 and l_2 are not located in the study area, three sub-routes $\{l_3, l_4, l_5, l_6, l_7\}$, $\{l_3, l_4, l_5, l_6\}$, and $\{l_4, l_5, l_6, l_7\}$ can be derived from the remaining part with the condition $q \geq 4$.

As a result, we can overcome the sparseness under different traffic patterns. Indeed, the sample size increases by at least one order of magnitude for each pattern (compare the 2nd and 4th columns of Table I). Further, the frequency of raw experiential routes can also be improved (see the 3rd and 5th columns). It is now evident that the updated database is highly informative as indicated in Table I.

B. Collaborative Preference Discovery (CPD)

In practice, the route discovery for an O-D pair may not sufficiently recommend the most preferable route. Therefore, the cluster-to-cluster retrieval method can be used to capture the top-k routes from the ERDB for a specific query traffic pattern.

In Algorithm 1, set several parameters and an O-D pair as fundamental inputs, the first step guarantees that we can derive two road segment sets within a prescribed t_c. The road segment set c_s (source cluster) satisfies that the travel time from the source to the specific l isn't longer than the threshold t_c, i.e., $c_s = \{l \mid t_s \leq t_c\}$, and road segment in c_d (destination cluster) is required to satisfy that the travel time from l to the destination isn't longer than t_c, i.e., $c_d = \{l \mid t_d \leq t_c\}$.

Algorithm 1 CPD Algorithm

Input:rank-threshold k, t_c, max $\{t_c\}$, source, destination
output: R_q, R_t
1: $c_s = \{l \mid t_s \leq t_c\}$, $c_d = \{l \mid t_d \leq t_c\}$
2: $Route = ERDB\,(c_s \to c_d)$
3: $C_{route} = f_{route} \cdot \log\left(\frac{2\max\{t_c\}+1}{t_s+t_d+1}\right), route \in Route$
//the collaborative degree of each extracted route
4: **if** $num\,\{Route\} \geq k$ **then**
5: $\left\{R_q, R_t\right\} \Leftarrow rank\,(Route, k)$
//top-k routes R_q are selected with respect to C_{route}
6: **else**
7: $\left\{R_q, R_t\right\} \Leftarrow rank\,(Route, num\,\{Route\})$
8: **end if**

Via the ERDB, we capture some experiential routes from c_s to c_d directly, and the retrieved candidate routes are stored into set $Route$. For $route \in Route$, C_{route} calculated by Equation (2) can be a criteria to rank the candidate routes. If the total number of candidate routes is more than k, we extract the candidate routes with top-k largest C_{route} values. Otherwise, while the candidate routes are less than k, the candidate routes are all remained. Notice that R_q denotes the query experiential route and R_t refers to the transfer route. As a result, we can derive R_q and R_t.

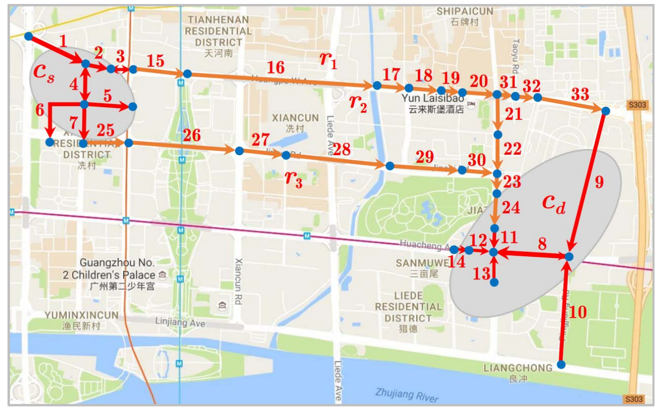

Fig. 6. Scenario of cluster-to-cluster retrieval.

TABLE II
INFORMATION ON THE DERIVED TOP-3 ROUTES

R_q	r_1	r_2	r_3
R_t in c_s	----	$\{l_1, l_4, l_7\}$	$\{l_1, l_2, l_3, l_{15}\}$
R_t in c_d	----	$\{l_{11}, l_8\}$	$\{l_{11}, l_8\}$
f_{route}	4	25	26
t_s	0	149.14	178.97
t_d	0	149.18	149.18
C_{route}	26.32	21.98	20.39

To better understand the Algorithm 1, a simplified actual scenario during the period of evening peak hours on weekdays is shown in Fig. 6. Given the source l_1 and the destination l_8, assumed that the rank-threshold $k = 3$, $t_c = 180$, and $\max\{t_c\} = 360$. From c_s to c_d, top-3 routes are captured and can be written as follows:

$$r_1 = \{l_1, l_2, l_3, l_{15}, \cdots, l_{20}, l_{31}, l_{32}, l_{33}, l_9, l_8\},$$
$$r_2 = \{l_{16}, l_{17}, \cdots, l_{23}, l_{24}\}, r_3 = \{l_{25}, l_{26}, \cdots, l_{30}, l_{23}, l_{24}\}.$$

In this case, r_1 is a complete route and there is no transfer route to connect either l_1 or l_8, however, r_3 has two transfer routes $\{l_1, l_4, l_7\}$, $\{l_{11}, l_8\}$ which correspond to c_s and c_d respectively. In Table II, the observations f_{route}, t_s, t_d indicate that C_{route} can enable a balance between frequency and transfer times. Even though f_{route} of r_2 and r_3 are much higher than that of r_1. The collaborative degree of r_1 is slightly higher than those of r_2 and r_3, and this implies that shorter transfer times tend to contribute a higher C_{route} value. If we only consider route frequency to select the top-2 routes, r_1 may not be an alternative due to its low frequency.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

C. Intelligent Driver Network Generation (IDNG)

The above Algorithm 1 can efficiently capture the top-k routes from the period-based ERDB. As an iterative procedure, these R_q and R_t derived from Algorithm 1 will be treated as fundamental inputs for the IDNG algorithm (i.e., Algorithm 2).

For $l \in R_q$, the road segments along R_q can be used to generate several extended clusters. Given an extended cluster $c_e = \{l | t_e \leq t_c\}$, the steps 3-9 retrieve $\left\{R_e^{(1)}, R_t\right\}$ where $R_e^{(1)}$ is extended experiential route and R_t is transfer route. Note that, for simplification, we store the results of all road segments satisfy $l \in R_q$ into the set $\left\{R_e^{(1)}, R_t\right\}$.

In the crucial key steps 11-13, only if a road segment l satisfies $l \in R_e^{(j)}$ (i.e., the road segment l is located on the jth R_e), $l \notin c_s$ (i.e., the road segment l is not located in the source cluster), $l \notin c_e$ (i.e., the road segment l is not located in the previous $j-1$th c_e), and $l \notin c_d$ (i.e., the road segment l is not located in the destination cluster), we execute a while loop to repeat lines 2-9, where the notation $\left\{R_e^{(1)}, R_t\right\}$ will change to the jth $\left\{R_e^{(j)}, R_t\right\}$. And $\left\{R_e^{(j)}, R_t\right\}$ is also a set which contains all c_e-derived R_e and R_t.

In the last step, as we have derived several experiential routes and their transfer routes, the IDN for a specific source and destination is comprised of $R_q, R_e^{(1)}, \cdots, R_e^{(m)}$ and their R_t where m in $R_e^{(m)}$ represents the total number of iterations in steps 11-13. Moreover, the IDN is structured as $G = (V, E)$ where V and E represent the intersections and road segments of all routes $R_q, R_e^{(1)}, \cdots, R_e^{(m)}, R_t$ in step 14.

Algorithm 2 IDNG Algorithm

Input:rank-threshold k, t_c, $\max\{t_c\}, j = 1$, R_q, R_t
output: $G = (V, E)$ //return an IDN
1: **for** $l \in R_q$ **do**
2: $c_e = \{l | t_e \leq t_c\}$
3: $Route = ERDB\,(c_e \rightarrow c_d)$
4: $C_{route} = f_{route} \cdot \log\left(\frac{2\max\{t_c\}+1}{t_e+t_d+1}\right), route \in Route$
5: **if** $num\,\{Route\} \geq k$ **then**
6: $\left\{R_e^{(1)}, R_t\right\} \Leftarrow rank\,(Route, k)$
7: **else**
8: $\left\{R_e^{(1)}, R_t\right\} \Leftarrow rank\,(Route, num\,\{Route\})$
9: **end if**
10: **end for**
11: **while** $\left\{l \middle| l \in R_e^{(j)} \& l \notin c_s, c_e, c_d\right\} \neq \emptyset$ **do**
12: Lines 2-9; $j = j + 1$
13: **end while**
14: $G = \left\{R_q, R_e^{(1)}, \cdots, R_e^{(m)}, R_t\right\}$

In Algorithm 2, when we select $l \in R_e$, overlapping may appear between neighboring clusters. Regardless of whether an overlapping road segment in a previous cluster is selected or not, that overlap will not be taken into account in the next cluster. Only if the derived route yields a higher C_{route} than those starting with the overlapping road segments do we capture the route as a candidate for further IDN generation. In this sense, an infinite loop of crucial key steps 11-13 can be prevented.

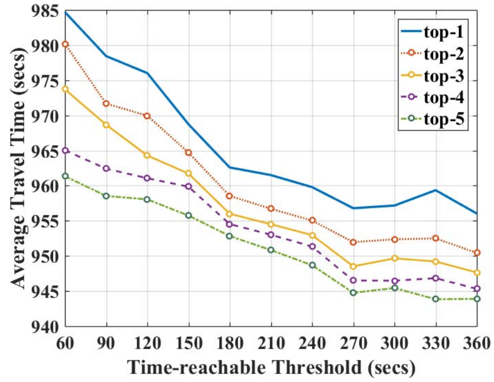

Fig. 7. Average travel time for various time-reachable thresholds and rank-thresholds.

V. Empirical Studies

In this section, relying on the ERDB constructed from taxis' trajectories, the 1500 random O-D pairs in our experiment are generated during weekdays include three traffic patterns, i.e., morning peak hours, off-peak hours, and evening peak hours. It is worthy to note that these O-D pairs are restricted by following. For an O-D pair, (1) there exists at least one experiential route from the source cluster to the destination cluster within the minimum time-reachable threshold of 60s, and (2) the minimum length of the routes is longer than 4,000 meters. Here, each traffic pattern contains 500 random O-D pairs separately. Then, we select the sensitivity-driven parameters t_c and k for the above algorithms, and compare the experiment performance among the proposed method and the conventional methods.

A. Sensitivity Analysis

Given a series of different time-reachable thresholds (i.e., 60, 90, 120, 150, 180, 210, 240, 270, 300, 330, and 360 seconds) and a series of the rank-thresholds (i.e., $k = 1, 2, 3, 4, 5$), using the proposed method, we can derive five curves. As plotted in Fig. 7, each curve represents the average travel time, which can be regarded as a function of time-reachable thresholds. From top to bottom, the five curves refer to the top-1, top-2, top-3, top-4, and top-5 routes, respectively. We can find that the average travel time tends to decrease as k increases from 1 to 5, and also decrease as the time-reachable threshold increases from 60s to 360s.

From the set of all parameter-varying curves in Fig. 7, we attempt to extract the optimal t_c and k. Representing the average travel time as a matrix with 11 rows and 5 columns, we can determine these two optimal parameters. The matrix of the average travel time is expressed as follows,

$$T = \begin{bmatrix} 984.70 & 980.19 & \cdots & 961.34 \\ 978.44 & 971.67 & \cdots & 958.54 \\ \vdots & \vdots & \ddots & \vdots \\ 956.03 & 950.44 & \cdots & 943.89 \end{bmatrix}_{11\times 5}$$

where the variance of each row is [97.11, 60.90, 51.64, 24.10, 14.52, 16.55, 17.57, 22.87, 22.52, 35.44, 23.10]^T; explicitly,

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III
PERFORMANCE MEASURE WITH RESPECT TO EFFICIENCY, RELIABILITY AND EFFECTIVENESS

Measures	Patterns	Proposed method	Fastest-path method	Shortest-path method
Connection	WDMP	25	990	1025
	WDOP	27	1108	1132
	WDEP	21	993	1063
Variance	WDMP	2454	2377	2676
	WDOP	1875	1825	2176
	WDEP	3588	3261	3994
Frequency	WDMP	3479	3185	3189
	WDOP	27100	24825	25171
	WDEP	2931	2678	2876
Travel time	WDMP	932	890	962
	WDOP	905	869	951
	WDEP	1109	1034	1149
Route length	WDMP	5893	5853	5601
	WDOP	6213	6142	5907
	WDEP	5994	5931	5655

[a]Connection refers to the connection scale (i.e., the number of road segments), variance is the variance of the travel time, frequency is the route frequency calculated by Equation (5).
[b]WDMP: weekday morning peak hours; WDOP: weekday off-peak hours; WDEP: weekday evening peak hours.

when the time-reachable threshold is 180s, we obtain the smallest variance (i.e., 14.52), which implies that the threshold 180s will recommend the most reliable routes under each of the five k values. Then, we extract 5th row from matrix T, i.e.,

$$T(5,:) = [962.59, 958.55, 956.00, 954.53, 952.83]$$

where the difference between top-1 and top-2 is 4.04, between top-2 and top-3 is 2.55, between top-3 and top-4 is 1.47, and between top-4 and top-5 is 1.60. Then, we can obtain the smallest decrease appears between top-3 and top-4, and thus, the optimal rank-threshold is $k = 3$.

B. Efficiency, Reliability and Effectiveness

As mentioned above, the optimal rank-threshold is $k = 3$ and the optimal time-reachable threshold is $t_c = 180$. With these optimal parameters, we present the desirability of the proposed method from three aspects (i.e., efficiency, reliability, and effectiveness) in the following empirical studies.

In terms of efficiency, we have introduced a connection scale as the quantity to measure the network scale. Given a source and a destination, the fastest-path and shortest-path methods computed by the Dijkstra's algorithm tend to traverse most road segments globally in an urban road network during the route planning, where every turning at an intersection will be chosen if the upstream road segment is traversed. As indicated in Table III, the connection scale of the IDN is significantly smaller than the conventional methods; indeed, the IDN obtains less than 3% connections of the conventional methods. This is because that, rather than traversing all feasible road segments and all turnings of each intersection, an IDN comprised of experiential routes and their transfer routes tends to avoid redundancy. Thus, we regard the proposed method as an efficient one for route recommendation. The benefit of suggesting a small-scale road network in our proposed method tends to efficiently execute the Dijkstra's algorithm while the fastest-path and shortest-path methods are inferior in this aspect.

In addition, the efficiency can be further discussed on the basis of algorithm run-times. As exemplified by the evening peak hours, the run-times of conventional methods is about 1ms in average for each query of 500 random O-D pairs (in detail, 1.03ms for the fastest-path method and 0.91ms for the shortest-path method), and the run-time of proposed method is 13.22ms. The above run-times are yielded after reading experiential routes from database into RAM, this means that the run-times of the proposed method is actually the total computing time of Algorithm 1, Algorithm 2, and Dijkstra's algorithm for finding the fastest route in an IDN.

In the real-world traffic, taxi drivers' route choices are generally more or less inflexible, because the accurate estimate of future traffic status is impossible naturally, even though taxi drivers accumulate their experience from daily driving. Whereas, the traffic status can be approximately perceived by taxi drivers, and it is evident that their experience is reliable and applicable, therefore the experienced taxi drivers can make their route decisions appropriately.

In terms of reliability and effectiveness, as shown in Table III and Fig. 8, (1) with respect to the variance of travel time, the routes recommended by the proposed method obtain less variance than the shortest-path method; however, it is not evident that the proposed method outperforms the fastest-path method in time reliability. (2) The routes recommended by the proposed method are traversed more frequently by taxis than those of the conventional methods. (3) The travel time and length of routes recommended by the proposed method are close to the shortest travel time and shortest route length, respectively, although the proposed method is not explicitly designed to recommend the fastest or shortest routes globally.

Since the experiential routes can provide steady routable information, the routes recommended by the proposed method are reliable and traversed more frequently by taxis. Intuitively, taxi drivers tend to choose a time-reliable route depending on the traffic patterns for a given O-D pair, and that route satisfies the drivers' time requirements for profitability. Here, the prior knowledge can be interpreted as latent factors related to route choice, such as comfort, convenience (e.g., unreasonable road designs) and safety (e.g., traffic violation and accident) of driving. Notice that the overall frequency of fastest routes is significantly lower than the other methods especially during the evening peak hours on weekdays, this may be the fact that the period with a rather congested traffic causes few taxi drivers to play tricks and choose route containing illegal travel, thereby, taking less travel time. And as a result of lacking samples, the fastest route generally in this manner may have smaller variance.

VI. FURTHER DISCUSSIONS

Essentially, although the performance has been measured to compare our proposed method with the conventional methods, we intend to further introduce a parallel comparison between the most popular route and the top-k routes which is

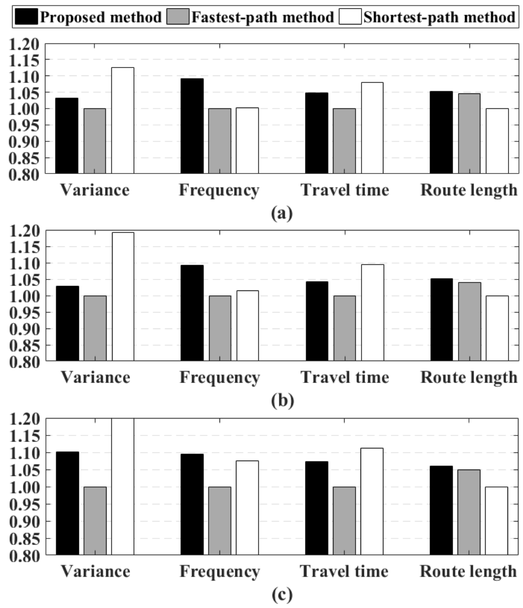

Fig. 8. Normalized performance measures in terms of reliability and effectiveness. (a) The morning peak hours on weekdays. (b) The off-peak hours on weekdays. (c) The evening peak hours on weekdays.

TABLE IV

COMPARISON RESULTS AMONG MPR, IMPR, AND THE TOP-K ROUTES

Measures	Patterns	The top-k routes	MPR	IMPR
Connection	WDMP	25	17	22
	WDOP	27	19	24
	WDEP	21	17	21
Variance	WDMP	2454	2490	2465
	WDOP	1875	1852	1868
	WDEP	3588	3732	3657
Frequency	WDMP	3479	3448	3471
	WDOP	27100	28625	27287
	WDEP	2931	2979	2936
Travel time	WDMP	932	958	946
	WDOP	905	923	951
	WDEP	1109	1133	1118
Route length	WDMP	5893	5980	5961
	WDOP	6213	6223	5907
	WDEP	5994	6038	6019

[a]Connection refers to the connection scale (i.e., the number of road segments), variance is the variance of the travel time, frequency is the route frequency calculated by Equation (5).

[b]WDMP: weekday morning peak hours; WDOP: weekday off-peak hours; WDEP: weekday evening peak hours.

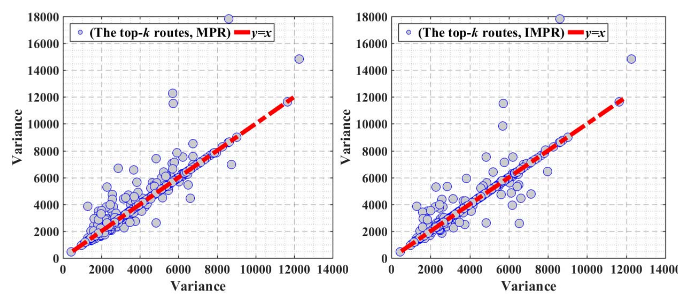

Fig. 9. Variance differences during evening peak hours on weekdays.

controlled by the collaborative degree. As noted in [8]–[11], the most popular route, and sometimes the top-k frequent routes, are typically considered during experiential route recommendation. In the cluster-to-cluster retrieval scenario, it is also instructive to compare the most popular route with the top-k routes to demonstrate that the routable information can be formulated and explained by the collaborative degree.

To be precise, the most popular route recommendations are classified into two experiment categorizes. The first is that, given the time-reachable threshold, we directly capture the most popular route between source cluster and destination cluster, and we briefly call it as MPR (i.e., most popular route). The second is that, according to the route frequency (two transfer times t_s and t_d are no longer considered), the top-1 route is captured iteratively, thus, we briefly call it as IMPR (i.e., iterative-MPR, the CPD and IDNG algorithms are utilized to generate an IDN).

Specifically, the differences among MPR, IMPR and the top-k routes lie in the rank-threshold k, the ranking measure (C_{route} or route frequency), and an iterative procedure is used or not. Given the same $t_c = 180$, the comparison results from Table IV indicate that (1) MPR, IMPR, and the top-k routes share approximately equal route frequency, travel time and route length, (2) the top-k routes imply a slightly larger connection scale than IMPR, and (3) with respect to time variance during evening peak hours, the routes recommended by the top-k routes method are more reliable than MPR and IMPR (depicted in Fig. 9).

Therefore, we can conclude that (1) the most popular route may not correspond to the most reliable route, this is because that the frequency of routes between two clusters are only considered while their transfer times are ignored, and (2) the collaborative degree which takes into account both route frequency and transfer times in the top-k routes tends to advise more reliable routes. To some extent, the collaborative degree, a ranking scheme considering both popularity and consistency, can be useful to express the routable information in the cluster-to-cluster retrieval.

VII. CONCLUSION

In big data era, discovering knowledge such as experience and preference from human trajectory for better travelling is an interesting topic. Thus, how to harness the large-scale trajectory data and unlock their power is rather meaningful. The route planning, as a classical problem, can also be accomplished by leveraging experience of taxi drivers' route decisions.

To this end, this study presents a collaborative route planning method with the latent knowledge from experienced drivers in urban areas. In general, the proposed method makes the route planning can be solved by a query-like solution, rather than the traditional graph-labelling method. Given an O-D pair, a rank-threshold k, and a time-reachable threshold t_c, the CPD and IDNG algorithms can be used to capture several experiential routes and generate an IDN.

Empirical studies demonstrate that our proposed method performs favorably in terms of efficiency, reliability, and effectiveness while comparing with the conventional fastest-path and shortest-path methods. Furthermore, the routes recommended by the proposed method tend to better express the latent experience and preferences of taxi drivers. Last but not least, with the proposed threshold free method in selecting a suitable t_c, the cluster-to-cluster retrieval based method can address the existing sparseness issue to some extent.

There also exist some limitations and requirements for the proposed method. First, due to the sparseness issue, the experiment area must have enough taxis and taxis' trajectory data. Second, if we apply the proposed method in a large-scale road network, the time-reachable threshold should be highly correlated with the distance of a given O-D pair (i.e., a distant O-D pair should set a larger t_c, while a near O-D pair corresponds to a smaller t_c).

REFERENCES

[1] S. Çolak, A. Lima, and M. C. González, "Understanding congested travel in urban areas," *Nature Commun.*, vol. 7, Mar. 2016, Art. no. 10793.

[2] Y. Zheng and X. Zhou, Eds., *Computing With Spatial Trajectories*. New York, NY, USA: Springer-Verlag, 2011.

[3] Y. Zheng, "Trajectory data mining: An overview," *ACM Trans. Intell. Syst. Technol.*, vol. 6, no. 3, pp. 1–41, May 2015.

[4] J. Yuan *et al.*, "T-drive: Driving directions based on taxi trajectories," in *Proc. ACM GIS*, 2010, pp. 99–108.

[5] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "T-drive: Enhancing driving directions with taxi drivers' intelligence," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 220–232, Jan. 2013.

[6] J. Yuan, Y. Zheng, X. Xie, and G. Sun, "Driving with knowledge from the physical world," in *Proc. ACM SIGKDD*, 2011, pp. 316–324.

[7] J. Zhang, W. Meng, Q. Liu, H. Jiang, Y. Feng, and G. Wang, "Efficient vehicles path planning algorithm based on taxi GPS big data," *Optik-Int. J. Light Electron Opt.*, vol. 127, no. 5, pp. 2579–2585, 2016.

[8] Z. Chen, H. T. Shen, and X. Zhou, "Discovering popular routes from trajectories," in *Proc. 27th IEEE Int. Conf. Data Eng.*, Apr. 2011, pp. 900–911.

[9] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Constructing popular routes from uncertain trajectories," in *Proc. ACM SIGKDD*, 2012, pp. 195–203.

[10] W. Luo, H. Tan, L. Chen, and L. M. Ni, "Finding time period-based most frequent path in big trajectory data," in *Proc. ACM SIGMOD*, 2013, pp. 713–724.

[11] L.-Y. Wei, Y. Zheng, and W.-C. Peng, "Mining popular routes from social media," in *Multimedia Data Mining and Analytics*. Cham, Switzerland: Springer, 2015, pp. 93–116.

[12] J.-H. Hu, Z. Huang, and J. Deng, "A hierarchical path planning method using the experience of taxi drivers," *Procedia-Social Behav. Sci.*, vol. 96, no. 6, pp. 1898–1909, Nov. 2013.

[13] Q. Li, Z. Zeng, T. Zhang, J. Li, and Z. Wu, "Path-finding through flexible hierarchical road networks: An experiential approach using taxi trajectory data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 13, no. 1, pp. 110–119, 2011.

[14] J. Zheng and L. M. Ni, "Modeling heterogeneous routing decisions in trajectories for driving experience learning," in *Proc. 16th ACM Int. Conf. Ubiquitous Comput.*, 2014, pp. 951–961.

[15] L.-Y. Wei, K.-P. Chang, and W.-C. Peng, "Discovering pattern-aware routes from trajectories," *Distrib. Parallel Databases*, vol. 33, no. 2, pp. 201–226, Dec. 2013.

[16] L. Zhuang, J. Gong, Z. He, and F. Xu, "Framework of experienced route planning based on taxis' GPS data," in *Proc. 15th IEEE ITSC*, Sep. 2012, pp. 1026–1031.

[17] H. Bast *et al.* (2015). "Route planning in transportation networks." [Online]. Available: http://arXiv:1504.05140

[18] H. Wang *et al.*, "R3: A real-time route recommendation system," *Proc. VLDB Endowment*, vol. 7, no. 13, pp. 1549–1552, Aug. 2014.

[19] N. J. Yuan, Y. Zheng, L. Zhang, and X. Xie, "T-Finder: A recommender system for finding passengers and vacant taxis," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 10, pp. 2390–2403, Oct. 2013.

[20] X. Xu, J. Zhou, Y. Liu, Z. Xu, and X. Zhao, "Taxi-RS: Taxi-hunting recommendation system based on taxi GPS data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 1716–1727, Aug. 2015.

[21] D. Zhang *et al.*, "Understanding taxi service strategies from taxi GPS traces," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 123–135, Feb. 2015.

[22] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numer. Math.*, vol. 1, pp. 269–271, Dec. 1959.

[23] Y. Donytsher, B. Galon, and Y. Kanza, "Storing routes in socio-spatial networks and supporting social-based route recommendation," in *Proc. 3rd ACM SIGSPATIAL Int. Workshop Location-Based Social Netw.*, 2011, pp. 49–56.

[24] C. Chen, D. Zhang, B. Guo, X. Ma, G. Pan, and Z. Wu, "TripPlanner: Personalized trip planning leveraging heterogeneous crowdsourced digital footprints," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1259–1273, Jun. 2015.

[25] H. Su, K. Zheng, J. Huang, H. Jeung, L. Chen, and X. Zhou, "CrowdPlanner: A crowd-based route recommendation system," in *Proc. 30th IEEE Int. Conf. Data Eng.*, Mar. 2014, pp. 1144–1155.

[26] B. Zheng, H. Su, K. Zheng, and X. Zhou, "Landmark-based route recommendation with crowd intelligence," *Data Sci. Eng.*, vol. 1, no. 2, pp. 86–100, Jun. 2016.

[27] S. Liu and Q. Qu, "Dynamic collective routing using crowdsourcing data," *Transp. Res. B, Methodol.*, vol. 93, pp. 450–469, Nov. 2016.

[28] W. T. Hsu, Y. T. Wen, L. Y. Wei, and W. C. Peng, "Skyline travel routes: Exploring skyline for trip planning," in *Proc. 15th IEEE MDM*, Jul. 2014, pp. 31–36.

[29] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, "Travel route recommendation using geotagged photos," *Knowl. Inf. Syst.*, vol. 37, no. 1, pp. 37–60, Oct. 2013.

[30] L. Liu, J. Xu, S. S. Liao, and H. Chen, "A real-time personalized route recommendation system for self-drive tourists based on vehicle to vehicle communication," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3409–3417, Jun. 2014.

[31] C.-Y. Tsai and B.-H. Lai, "A location-item-time sequential pattern mining algorithm for route recommendation," *Knowl.-Based Syst.*, vol. 73, pp. 97–110, Jan. 2015.

[32] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Comput.*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[33] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[34] Y. Chen, M. G. H. Bell, and K. Bogenberger, "Reliable pretrip multipath planning and dynamic adaptation for a centralized road navigation system," *IEEE Trans. Intell. Transp. Syst.*, vol. 8, no. 1, pp. 14–20, Mar. 2007.

[35] M. Kobitzsch, D. Schieferdecker, and M. Radermacher, "Evolution and evaluation of the penalty method for alternative routes," in *Proc. ATMOS*, 2013, pp. 1–14.

[36] A. Efentakis, D. Thedorakis, and D. Pfoser, "Crowdsourcing computing resources for shortest-path computation," in *Proc. ACM SIGSPATIAL GIS*, 2012, pp. 434–437.

[37] M. Baum, J. Dibbelt, T. Pajor, and D. Wagner, "Dynamic time-dependent route planning in road networks with user preferences," in *Experimental Algorithms*, A. Goldberg and A. Kulikov, Eds. Cham, Germany: Springer, 2016.

[38] G. V. Batz, R. Geisberger, P. Sanders, and C. Vetter, "Minimum time-dependent travel times with contraction hierarchies," *J. Experim. Algorithms*, vol. 18, Dec. 2013, Art. no. 1.4.

[39] A. Paraskevopoulos and C. Zaroliagis, "Improved alternative route planning," in *Proc. ATMOS*, 2013, pp. 108–122.

[40] M. Kobitzsch, M. Radermacher, and D. Schieferdecker, "Evolution and evaluation of the penalty method for alternative graphs," in *Proc. ATMOS*, 2013, pp. 94–107.

[41] M. Renz, M. Schubert, and H.-P. Kriegel, "Route skyline queries: A multi-preference path planning approach," in *Proc. 29th IEEE Int. Conf. Data Eng.*, Mar. 2010, pp. 261–272.

[42] B. Yang, C. Guo, C. S. Jensen, M. Kaul, and S. Shang, "Stochastic skyline route planning under time-varying uncertainty," in *Proc. 30th IEEE Int. Conf. Data Eng.*, Mar. 2014, pp. 136–147.

[43] J. Dai, B. Yang, C. Guo, and Z. Ding, "Personalized route recommendation using big trajectory data," in *Proc. 31st IEEE Int. Conf. Data Eng.*, Apr. 2015, pp. 543–554.

[44] B. Yang, C. Guo, Y. Ma, and C. S. Jensen, "Toward personalized, context-aware routing," *VLDB J.*, vol. 24, no. 2, pp. 297–318, Apr. 2015.

[45] D. Delling, A. V. Goldberg, M. Goldszmidt, J. Krumm, K. Talwar, and R. F. Werneck, "Navigation made personal: Inferring driving preferences from GPS traces," in *Proc. 23rd ACM SIGSPATIAL GIS*, 2015, p. 31.

[46] S. Funke, S. Laue, and S. Storandt, "Deducing individual driving preferences for user-aware navigation," in *Proc. 24th ACM SIGSPATIAL GIS*, 2016, p. 14.

Zhaocheng He received the B.S. and Ph.D. degrees from Sun Yat-Sen University, Guangzhou, China, in 2000 and 2005, respectively.

He is currently a Professor with the Guangdong Provincial Key Laboratory of Intelligent Transportation Systems (ITS), and the Research Center of ITS, Sun Yat-Sen University. His research interests include traffic flow dynamics and simulation, traffic data mining, and intelligent transportation systems.

Kaiying Chen received the B.S. degree from Sun Yat-Sen University, Guangzhou, China, in 2016, where he is currently pursuing the M.S. degree from Sun Yat-Sen University. His research interests include traffic data mining and intelligent transportation systems.

Xinyu Chen received the B.S. degree from Guangzhou University, Guangzhou, China, in 2016.

He is currently pursuing the M.S. degree from Sun Yat-Sen University, Guangzhou. His research interests include traffic data mining and intelligent transportation systems.