This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

A Shared Bus Profiling Scheme for Smart Cities Based on Heterogeneous Mobile Crowdsourced Data

Xiangjie Kong, *Senior Member, IEEE,* Feng Xia, *Senior Member, IEEE,* Jianxin Li, *Member, IEEE,* Mingliang Hou, Menglin Li, Yong Xiang, *Senior Member, IEEE*

***Abstract*—Mobile crowdsourcing, as an effective and crucial technique of Industrial Internet of Things, is enabling smart city initiatives in the real world. It aims at incorporating the intelligence of dynamic crowds to collect and compute decentralized ubiquitous sensing data which can be used to solve major urbanization problems such as traffic congestion. The shared bus, as a neotype transportation mode, aims at improving the resource utilization rate and maintaining advantages of convenience and economy. In this paper, we provide a scheme to profile shared buses through heterogeneous mobile crowdsourced data (TRProfiling). First, we design an MCS-based shared bus data generation and collection solution to overcome the above data scarcity issue. Then we propose a Travel Profiling (TP) to profile resident travel and design a method called Multi-Constraint Evolution Algorithm (MCEA) to optimize the routes. Experimental results demonstrate that TRProfiling has an excellent performance in satisfying passengers' travel requirements.**

***Index Terms*—Industrial Internet of Things, travel profiling, mobile crowdsensing, route planning, shared buses.**

I. Introduction

RECENT years have witnessed a proliferation of Industrial Internet of Things (IIoT) techniques [1], including cyberphysical systems (CPS) [2], Internet of Things, automation [3], cloud computing [4], Internet of services [5], wireless technologies, etc. One important application scenario of IIOT is the smart city which aims at improving the public services in urban environments and dealing with problems in urbanization such as traffic congestion, energy consumption, and environmental pollution. The primary challenge in smart city consists of two aspects. One is how to collect and capture the vast amounts of dynamic data effectively in a pervasive environments. Another is how to analyze these multi-source heterogeneous spatial temporal data and then preciously construct profiling through data for specific tasks.

Mobile Crowdsourcing (MCS), as an effective and crucial technique of IIOT, devotes to connecting a plethora of mobile devices endowed with several sensing, actuation, and computing capabilities with the wireless network, thus providing decentralized ubiquitous services and applications in the context of a smart city [6]. Recently, a large number of research methods based on MCS have been performed and abundant applications are thus enabled [7]. In the field of intelligent transportation systems, MCS also provides new ideas for many issues such as route planning [8], human mobility pattern exploring [9], and traffic anomaly detection. It provides an excellent solution to the challenge above-mentioned in smart city. Fig. 1 displays the conceptualization process of using crowdsourcing to collect shared bus data.

In relatively large cities, traffic congestion during peak periods has turned into a daily routine for residents. The congestion is owing to the unreasonable resource allocation ultimately. Then in the case of a limited land area, confronted with the rapid growth in population and vehicles' number, how to relieve and eradicate the increasingly serious urban traffic conditions is plaguing relevant domain experts. Compared with crowded time-consuming buses, relatively expensive taxi and excessive fragmented trips of Online car-hailing services, the shared bus, which are emerging recently, as a neotype transportation mode, are benefitting from its convenience, economy, and instantaneity features. However, the above-mentioned advantages bring great resistance to the provision and promotion of shared bus services. The core of convenient,

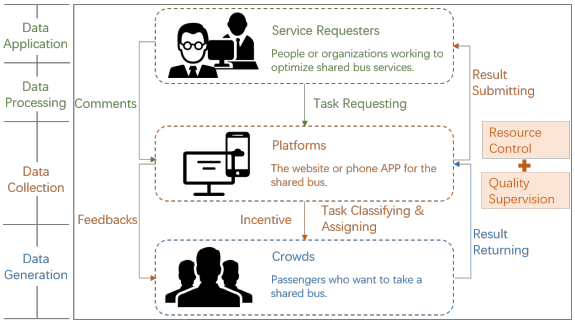


Fig. 1: The conceptualization process of using crowdsourcing to collect shared bus data.

This work was partially supported by the National Natural Science Foundation of China under Grant No. 61572106, the Dalian Science and Technology Innovation Fund under Grant No. 2018J12GX048, the Fundamental Research Funds for the Central Universities under Grant No. DUT18JC09 and the ARC Discovery Project under Grant No. DP160102114. (Corresponding author: Feng Xia.)

X. Kong and M. Hou are with School of Software, Dalian University of Technology, Dalian 116620, China (e-mail: xjkong@ieee.org; teemohold@outlook.com).

F. Xia is with School of Science, Engineering and Information Technology, Federation University Australia, Ballarat, VIC 3353, Australia (e-mail: f.xia@ieee.org).

J. Li and Y. Xiang are with the School of Information Technology, Deakin University, Burwood, VIC 3216, Australia (e-mail: jianxin.li@deakin.edu.au; yong.xiang@deakin.edu.au).

M. Li is with Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore (e-mail: cookies.s@outlook.com).

1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

instant bus services is to meet passenger demand dynamically, namely, dynamic route planning. The growth of passengers' number and stations' number greatly increases the complexity of dynamic route planning. This is the key technical reason why it is impossible to provide a true shared bus service (that is, to have all of the above advantages) and to promote it to the market. What's more, the immaturity of shared bus services inevitably results in the scarcity of shared bus data, and thus the dynamic route planning-oriented research is hard to push. The development of shared buses encounters bottlenecks that are urgent to be broken. The challenges of routing strategy of shared buses can be summarized as four parts. First is insufficient data. Route planning problems extremely rely on real-world data. However, there is little benchmark dataset that can be evaluated about shared buses. Second is insufficient researches. As a novel type of transportation, there is little research for shared buses routes planning to provide guidance. Third is high complexity. It's hard to characterize passengers' travel requirements accurately. Traditional methods only aim at maximizing carrying capacity which brings terrible travel experiences for passengers. Last is real-time requirements. Shared-bus services require real-time responses to calculate routes based on current traffic conditions.

Shared buses devote to improving resource utilization and provide a dynamic transportation mode. The conclusion that the core of shared bus service optimization lies in the dynamic integration of travel resources based on public demand can be drawn by us. In summary, the overall framework of this work is composed of three steps: Firstly, we collect the passenger's order data and the driver's GPS data through Futurefleet, a mobile phone shared bus APP. Secondly, after acquiring shared bus data, the crucial problem is how to understand and analyze resident travel requirements from multiple aspects and then construct accurate profiles for passengers. Profiling is an effective tool to delineate target users and reflect their needs. In this work, we propose a Travel Profiling (TP) to describe resident travel, and refine it into travel time, waiting time, seat utilization rate, delay tolerance, and loss tolerance. Finally, based on TP, we design a method called Multi-Constraint Evolution Algorithm (MCEA) to optimize the routes. The contributions of our work can be concluded as follows:

- We develop a scheme, TRProfiling, by merging shared bus data generation and collection, travel requirement description method (TP), and route optimization algorithm (MCEA), to profile shared buses and offer constructive suggestions for its development.
- We propose TP to portray resident travel requirements and refine it into travel time, waiting time, seat utilization rate, delay tolerance, and loss tolerance based on shared bus data analysis, and instantiate them by utilizing machine learning algorithms.
- We design a multi-constraint-based heuristic algorithm to generate the optimal route based on TP. The experimental results demonstrate that the optimal route can satisfy the passengers' requirements to the most degree.
- We conduct extensive experiments to verify the superiority of our method. The results show that our proposed scheme has an excellent performance in satisfying the passengers' travel needs.

The rest of this paper is organized as follows. In Section II, we review the related work about MCS and route planning methods, as well as their overlaps. Section III describes our proposed approach TRProfiling in detail. Following that, we conduct extensive comparative experiments to verify the effectiveness of our approach in Section IV. Finally, we conclude our work and further discuss the open issues in Section V.

II. Related Work

In this section, we mainly introduce the related work consisting of two parts: firstly, we review the current researches on MCS. And then, we introduce the related studies on route planning.

A. MCS

MCS has become a promising paradigm with the development of smartphone sensing and mobile social networking techniques. Compared to other sensing modalities, MCS has advantages of cross-space and large-scale [10]. MCS allows mobile phone users to share information such as traffic conditions and location information. Additionally, users can further upload the data to the cloud for large-scale sensing and community intelligence mining [11]. In particular, large amounts of MCS data provides a novel opportunity for research, which further can be applied into many other fields. These applications include the mobile social recommendation, environment monitoring, and traffic planning.

In the field of mobile social recommendation, Zheng *et al*. [12] measure the similarity among users and provide a personalized place recommendation service. Ye *et al*. [13] also develop a place recommendation service by exploring user-generated data in location-based social networks. Furthermore, they design a collaborative recommendation algorithm which consider the geographical influence on user check-in behaviors. In the area of ecological monitoring, a participatory noise mapping system is proposed [14].The authors of [15] and [16] study the short-term recomsumption behaviors and repeatable recommendation on user check-in dataset. The authors [14] use mobile phones to determine environmental noise level. Besides, mobile phones can also be used to gather information about trucks in order to measure air pollutions [17]. For traffic planning, Calabrese *et al*. [18] design a system which can report real-time urban dynamics by using traffic data. Based on crowd-powered data, Wolfson *et al*. [19] design a taxi ridesharing service called T-Share which can generate optimized ridesharing schedules. Li *et al* [20] study the problem of finding a destination place for a group of users and they design a framework which can be used to compute the exact query result and approximate query result for user.

B. Route Planning

Route planning is a hot research topic in the academic field, and it has important applications in many fields, such as self-driving, data forwarding, and GPS navigation. In the field of

1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

transportation, it usually means to find the best route given origin-destination pair (OD) under predefined conditions. The most famous and classical route planning problem is the Traveling Salesmen Problem (TSP), which is a well-known NP-hard problem in combinatorial optimization domain [21]–[23]. Many route planning problems can be categorized as TSP or its extensions including shared bus route planning problems.

To solve the traffic route planning problems, many scholars have conducted plentiful researches. Thomas *et al*. [24] present a system for individual trip planning, which incorporates future traffic hazards in routing. Wang *et al*. [25] develop an efficient indexing technique for route planning on timetable graphs called TTL. Based on the system, they further propose query algorithms that enable TTL to support three popular types of route planning queries. By carefully adapting node contraction, researchers are able to compute point-to-point queries on a continental network combined with cars, railroads, and flights [26]. Shang *et al* [27] solve the problem of efficient processing trajectory similarity joins by developing search space pruning techniques and make the parallel processing available.

C. Route Planning with MCS

With the prevalence of smart mobile devices, the scale of the crowdsensing system has rapidly increased, and a large number of MCS data can be readily collected. This crowd data can be used for traffic prediction, public transportation system design, as well as route planning.

Many studies have investigated route planning using crowdsensing data. For instance, B-Planner is a two-phase approach to exploring the bus route during night time [28]. In [28], the crowdsourced GPS data from taxis is first used to build a candidate bus stop set and then a bidirectional probability-based spreading algorithm is developed to generate candidate bus routes automatically. Yang et al. [29] define stochastic skyline routes mechanism considering multiple costs and time-dependent uncertainty based on crowdsourced GPS data from vehicles. Then they propose efficient algorithms to retrieve the best route for a given OD pair and start time. In [30], the bus passengersâĂŹ surrounding environmental context is utilized to estimate the bus arrival time, which is the primary information for most travelers, for different bus stations.

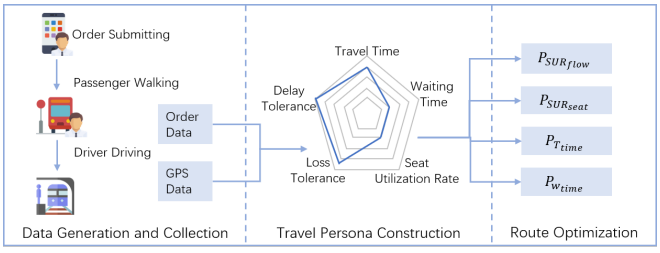


Fig. 2: Flow of TRProfiling.

III. TRPROFILING

Fig. 2 presents the framework of TRProfiling. It consists of three major components: 1) data generation and collection; 2) travel profiling construction; 3) route optimization. We describe them in detail in this section. The main notations used in TRProfiling can be found at Table I.

TABLE I: The main notations used in TRProfiling.

Notations	Descriptions
TP	Travel Profiling vector
T_{label_n}	Passenger's requirement label
S_{wt}	Waiting time of a route
S_{pass}	Total number of passengers of a route
N_{lp}	Number of passengers who leave the station
N_{wp}	Number of passengers of whole passengers
Z	Multi-cost of MCEA
P_{SUR}	Seat utilization rate of a route
F_m	Number of passengers of candidate station m
F^*	Number of passengers that has already got on the bus
$(s_1, s_2, ..., s_k)$	Set of all candidate stations
P_{T_time}	Total travel time of a route
T_m	Travel time to candidate station m
P_{W_time}	Waiting time of a route
W_m	Waiting time to candidate station m

A. Data Generation and Collection

Futurefleet is a mobile phone APP focused on providing shared bus services. In order to cope with the deficiency of shared bus data, we design it in partnership with an Internet company committed to enhancing travel experiences. Main participants of shared bus services are passengers and drivers. Therefore, Futurefleet is refined into driver client and passenger client for the sake of fully acquiring data.

A shared bus trip involves basic factors such as departure stations, arrival stations, departure time, and fare. The trips in the last mile scene that our work focuses on, is from residential areas to the nearest subway stations. Therefore, it's reasonable to regard arrival stations as known. For areas where shared buses are new, other unknown factors should be initialized according to passenger requirements. The APP attracts passengers to suggest departure stations and time, fare, and so on to participate in bus route customization. To ensure the authenticity of the data, passengers need to pay small amount deposit, which could then be refunded or as fare. We formulate the departure schedule, stations, and operation routes based on experience with reference to the above information. We initialize shared buses with fixed bus routes.

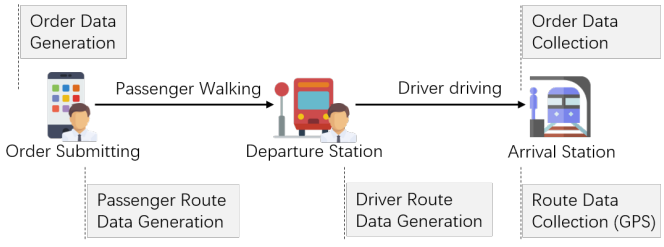


Fig. 3: Simplified representation of data generation and collection.

After initialization, shared buses are put into actual operation. Through Futurefleet, passengers could query real-time bus operational information, like locations and the number

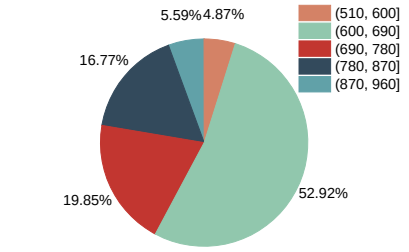
1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

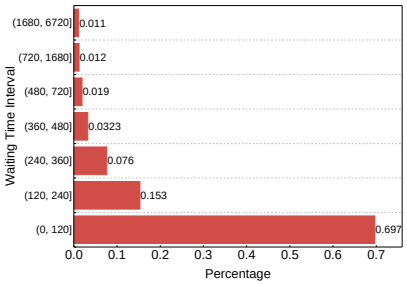
of vacant seats. They submit orders containing destinations and departure time in the APP. The nearest departure station would be assigned to the passenger. Within the scope of authority, passengers' real-time routes before boarding would be recorded. The formulated fixed shared bus routes are sent to drivers. Drivers follow the recommended routes unless an abnormal event is encountered, like accidents and bad weather. The APP records the change of routes in the form of GPS, which is employed to subsequent route optimization. The simplified representation of data generation and collection procedure is displayed in Fig. 3.

In this way, by submitting orders and going to departure stations, the information of passengers' order and route is generated separately. Drivers follow the published fixed routes or emergency diversion, and the relevant GPS data is thus recorded. The mobile terminal, that is, the APP, sends the above GPS information and order information to the cloud. After a simple integration, the collection of initial structured shared bus data is basically achieved.

In the follow-up study of this work, we focus on how to use shared bus data to analyze residents' travel requirements, that is, to profile residents' travel, and generate real and flexible shared bus routes that dynamically change with passenger demand.



(a) The distribution of travel time.



(b) The distribution of waiting time.

Fig. 4: The statistical analysis of passenger travel behaviors.

B. Travel Profiling Construction

We formulate Travel Profiling as a vector, which is shown in Equation. (1).

$$TP =< T_{label1}, T_{label2}, T_{label3}, ..., T_{labeln} > \quad (1)$$

where T_{labeln} is the label that reflects passengers' requirements. Labels contained in TP represent travel requirements from different aspects and this enables TP interpretability. TP could be obtained through statistical analysis of passenger travel behaviors. Next we present a detailed description of TP initialization (labels' refinement) and TP instantiation (labels' numeralization).

1) Travel Profiling Initialization: Since passengers prefer to arrive at destinations as quickly as possible, we statistically analyze the travel time and waiting time according to passengers' travel behaviors. Fig. 4(a) shows the distribution of travel time for passengers and are divided into five time intervals. It is obvious that the travel time of more than half of the passengers mainly centralizes in 10 minutes, which illustrates that the short distance travel is the focus of shared buses passengers. This puts the timing requirement with higher priority. Fig. 4(b) represents the distribution of the waiting time of passengers. From the figure, we can observe that nearly 70 percents of passengers prefer to wait less than 2 minutes at one station and no one accepts to wait more than eight minutes. We can summarize that passengers would tolerate waiting for a limited time range. When the waiting time is higher than the threshold, passengers would leave the station and choose the other transportation method. Therefore, we introduce two definitions, i.e., delay tolerance and loss tolerance to describe this scene. Besides, considering the company's profit, the Seat Utilization Rate (SUR) of each path is another factor that should be considered. We count the SUR of each route of seven days from Monday to Sunday, as shown in Table. II. After the above analysis, the labels of TP can be summarized as follows:

Travel Time The time that passengers spend on the shared bus is called as travel time. It is equal to the span between the boarding time and the time of getting off. Every passenger wants to reach the destination as fast as possible with the help of the travel time.

Waiting Time An essential factor to influence the travel experiences of passengers is the waiting time. The waiting time refers to the time passengers spending at one station waiting for the shared bus, which can be considered as the time from creating the order to checking the ticket.

Seat Utilization Rate To meet the demand for the profit of the company, we introduce the label of seat utilization rate, which controls whether the shared bus company can gain profit or not. The seat utilization rate can be determined by the passenger flow and the number of seats of the shared buses.

Delay Tolerance The waiting time is another essential element among the passengers' demands. We assume that every passenger has a maximal threshold for the waiting time. When the waiting time is greater than the threshold, passengers will leave the station. Delay tolerance is used to representing the threshold, and it is equal to the average value of the waiting time.

Loss Tolerance Shared buses will lose passengers if the waiting time for passengers is greater than the delay tolerance. The loss tolerance of a station can be obtained by the number of passenger loss divided by the total number of the station. Shared buses will yield unsatisfactory experiences for passengers if the loss tolerance is too large.

1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

TABLE II: The seat utilization of routes.

Routes \ Week	Mon	Tues	Wed	Thur	Fir	Sat	Sun
Route 1	0.465	0.456	0.471	0.455	0.479	0.14	0.09
Route 2	0.246	0.246	0.265	0.269	0.27	0.086	0.053
Route 3	0.535	0.526	0.544	0.529	0.55	0.165	0.106

2) Travel Profiling Instantiation: In order to get the input and constraints of the MCEA, a series of methods are designed to instantiate TP. Detailed descriptions of how to instantiate each label of TP are introduced as follows:

Travel Time Estimation Due to the terrible traffic condition caused by traffic jams, the travel time continuously varies in the different time intervals. According to the operation rules of shared buses, the travel time is counted every 15 minutes. Based on (TTM), the travel time between every two stations at a particular time can be obtained. The TTM is one of the inputs of the MCEA which can be used to calculate $P_{T_{time}}$ defined in Equation. (17).

Waiting Time Estimation The waiting time is equal to the difference between order creation time and the bus arrival time. Thus, the order creation time is predicted in order to calculate the waiting time. To predict the order creation time of each station, the AutoRegressive Integrated Moving Average model (ARIMA) is employed. Based on the TTM and the start time of the shared bus, the arrival time that the shared buses arrive at each station can be obtained. After that, the Waiting Time Matrix (WTM) which is used to calculate the $P_{W_{time}}$ in Equation. (18) can be calculated.

Seat Utilization Rate Analysis Seat utilization rate is the ratio of the actual passenger capacity of a route and the total number of seats on the shared bus. A bus should maximally utilize its capacity as well as make sure the bus is not overloaded. In order to calculate the actual passenger capacity, we employ machine learning algorithm to predict the distribution and volume of boarding passengers for each station during different time intervals. The input features include historical flow, time, distance, and week. After prediction, the Passenger Flow Matrice (PFM) can be formed to calculate the $P_{SUR_{flow}}$ in Equation. (15), as well as the seat utilization rate.

Delay Tolerance Analysis Delay tolerance is a constraint for the MCEA. It refers to the maximal time that passengers willing to wait. Based on the WTM, the waiting time of a route can be calculated. Delay Tolerance is equal to the average of the waiting time in Equation. (2)

$$D_t = \frac{S_{wt}}{S_{pass}} \tag{2}$$

where S_{wt} is the waiting time of a route that can be obtained from the WTM, and S_{pass} is the total number of passengers of a route which can be retrieved from the PFM. The delay tolerance of a route must be less than n based on the constraints (8).

Loss Tolerance Analysis Passengers will lose their patience if they wait too long. Loss tolerance is another constraint of a route which is used to measure the rate of passengers' drain. Before the shared bus reaches a station, the number of the passengers who leave the station can be counted. Therefore, the loss tolerance of the station can be calculated as Equation. (3)

$$L_t = \frac{N_{lp}}{N_{wp}} \tag{3}$$

where N_{lp} is the number of the passengers who leave the station, and N_{wp} is the number of whole passengers. The loss tolerance of a route is the average of the loss tolerance of all the stations on the route. And it must be less than m according to constrains (9).

C. Route Optimization

1) Problem Formulation: An optimal route taking into account multi-constraints can be modeled as a TSP problem which is a typical NP-hard problem in combinatorial optimization. To solve this problem, we model it as a completely connected graph $G(V, E)$ in an N-dimensional Euclidean space (N is the size of the stations), in which stations are the graph's vertices, paths are the graph's edges, and a path's multi-cost is the edge's length. Mathematically, the problem can be defined as given a set of n stations, the goal is to discover a route for these stations to minimizes Z which is the sum of the multi-cost of each path(i, j) on the route. The overall system parameters can be abstracted as follows:

$$min \quad Z = \gamma(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}x_{ij}) \tag{4}$$

$$c_{ij} = t_{ij} + w_{ij} \tag{5}$$

$$s.t. \begin{cases} \sum_{i=1}^{n} x_{ij} = 1, \forall j \in V & (6) \\ \sum_{j=1}^{n} x_{ij} = 1, \forall i \in V & (7) \\ \sum_{i\in S}\sum_{j\in S} x_{ij} \le |S| - 1, \forall S \subset V, 2 \le |S| \le n-1 & (8) \\ k < S_r \le 1 & (9) \\ D_t < n & (10) \\ L_t < m & (11) \\ x_{ij} \in \{0, 1\} & (12) \\ \gamma \in [1, \infty) & (13) \end{cases}$$

$$x_{ij} = \begin{cases} 1, & path(i,j) \quad is \quad on \quad the \quad route \\ 0, & path(i,j) \quad isn't \quad on \quad the \quad route \end{cases} \tag{14}$$

The objective function (2) minimizes the multi-cost c_{ij} of the route, which contains travel time t_{ij} and waiting time w_{ij} in the equation. 5. Constraints (4)(5) ensure every station on the route must be visited only once and constraints (6) is the underlying elimination constraints. Constraint (7) provides a minimal value for the seat utilization rate to ensure the profit of the company. Constraints (8)(9) ensure that loss tolerance and delay tolerance must be lower than m and n respectively, which are obtained from the statistical analysis. In the Constraint (10), x_{ij} is a binary variable judging whether a path(i, j) is on the route or not, and it is shown as equation.12. The γ in the contains (11) is the reciprocal of load factor, which is no less than 1.

1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

2) Multi-Constraint Evolution Algorithm: Our work aims to find the optimal route to minimize the multi-cost Z, which is defined in Equation. (4), and the optimal route satisfies the constraints that we get from the TP instantiation, i.e., delay tolerance, loss tolerance, and seat utilization rate. In order to generate the optimal route, we design a heuristic algorithm called MCEA. The fundamental idea of the MCEA is that given an OD pair, we select the next station based on the probability P and add that station to the visited set. We repeat this process until all the stations are visited. Then, a new route can be generated. The multi-cost Z can be updated if the multi-cost Z of the new route is less than the old one. The algorithm will converge until the Z is smaller than predefined threshold. And finally, the optimal route can be obtained. The probability P is computed by four parts, i.e., $P_{SUR_{flow}}$, $P_{SUR_{seat}}$, $P_{T_{time}}$, and $P_{W_{time}}$. The definitions of them are given below.

- P_{SUR}: P_{SUR} can influence the seat utilization rate of a route, which contains $P_{SUR_{flow}}$ and $P_{SUR_{seat}}$. $P_{SUR_{flow}}$ is used to measure the influence of passenger flow on the selected routes. Usually, the station which has more passengers can improve the seat utilization rate, and has a higher probability to be chosen as the next station. The $P_{SUR_{flow}}$ is computed by

$$P_{SUR_{flow}}\{s_m|(s_1, s_2...s_k)\} = \frac{F_m + F^*}{\sum_{i=1}^{k} F_i + F^* * k} \tag{15}$$

  here, F_m is the number of passengers of candidate station m, which can be retrieved from the PFM, and $(s_1, s_2...s_k)$ is the set of all candidate stations. F^* is the number of passengers that already get on in the previous stations, and it can be employed to avoid local minimum. $P_{SUR_{seat}}$ can measure the importance of the number of seats for the selected routes. To ensure passengers' travel experiences and safety, every passenger should have a seat, and the bus returns to the terminal as soon as it's full. Therefore, the bus should always move forward to the destination with the continuous improvement of the load factor. The $P_{SUR_{seat}}$ is computed by

$$P_{SUR_{seat}}\{s_m|(s_1, s_2...s_k)\} = \frac{\sum_{i=1}^{k} S_{i\_end} - S_{m\_end}}{\sum_{i=1}^{k} S_{i\_end} * (k-1)} \tag{16}$$

  where S_{m_end} is the distance from candidate station m to the end, which can be got from the TDM, so TDM is also an input of the MCEA.
- $P_{T_{time}}$: The total travel time of a route can be controlled by $P_{T_{time}}$. Usually, the station which has the shorter travel time are more likely to be chosen. The $P_{T_{time}}$ is computed by

$$P_{T_{time}}\{s_m|(s_1, s_2...s_k)\} = \frac{\sum_{i=1}^{k} T_i - T_m}{\sum_{i=1}^{k} T_i * (k-1)} \tag{17}$$

  where T_m is the travel time to candidate station m, which can be got from the TTM.
- $P_{W_{time}}$: $P_{W_{time}}$ can control the total waiting time of a route. If a station has a shorter waiting time, it has the higher probability to be chosen as the next station. The $P_{W_{time}}$ can be computed by

$$P_{W_{time}}\{s_m|(s_1, s_2...s_k)\} = \frac{\sum_{i=1}^{k} W_i - W_m}{\sum_{i=1}^{k} W_i * (k-1)} \tag{18}$$

  where W_m is the waiting time to candidate station m, which can be got from the WTM.

According to the probability theory, we assume X and Y are two independent random variables, and their probability functions are $X(z)$ and $Y(z)$. Q represents the sum of X and Y, and the probability function of Q is equal to the convolution of $X(z)$ and $Y(z)$ [31]. Therefore, the final probability P can be computed by

$$P = P_{SUR_{flow}} \bigodot P_{SUR_{seat}} \bigodot P_{T_{time}} \bigodot P_{W_{time}} \tag{19}$$

where P is the final probability of the MCEA to select the next station randomly. Algorithm 1 is the pseudocode of the MCEA. In Lines 5-13 of 1, the final probability P is calculated according to Equation (19) and then the next station s_i is selected according to the calculated probability. In Lines 14-20, the constraints (9)-(11) of the objective function (4) is computed, followed by updating the four elements of probability P in Equation (19).

IV. EXPERIMENTS

In order to demonstrate the effectiveness of our method, we conduct experiments of our proposed scheme TRProfiling. We first describe the details of shared bus data and data processing and then provide the evaluation of TRProfiling compared with other algorithms based on the obtained real-world shared bus datasets.

A. Data Description

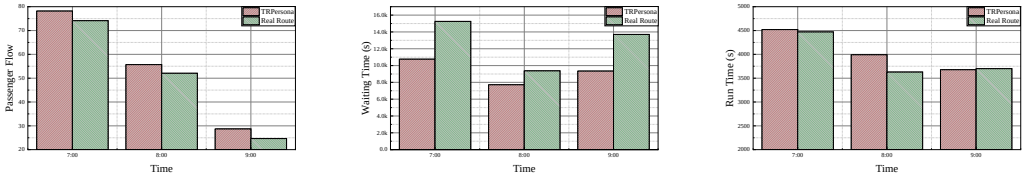
We cooperate with the Futurefleet and operate shared bus services in Yongkang city, Minhang Zone, Shanghai and spend more than six months to collect the datasets. Yongkang City is a large residential area and the distance from residential homes to the nearest stations is average 2KM. Every morning, a large number of residents need to commute by subway. Fig.5 shows the application area of the experiments. The dataset consists of order data, and GPS trajectory data from April 1st, 2017 to September 6th, 2017. The order data contains information about passengers' travel behaviors, e.g., order ID, passenger ID, station ID of passenger boarding, order creation date and time, passenger boarding date and time. It consists of about 50,000 records of passengers from 6:00 a.m. to 22:00 p.m., including payment in cash and canceled orders. The GPS trajectory data covers all stations of Yongkang City, Shanghai. The information includes bus IDs, longitude and latitude, recording time, last station ID, region code, bus driver information, azimuth, precision, etc. The detailed description of the dataset is shown in Table III and Table IV.

B. Data Preprocessing

During the data preprocessing stage, we preserve the data between 6:40 a.m. and 9:40 a.m., because we focus on

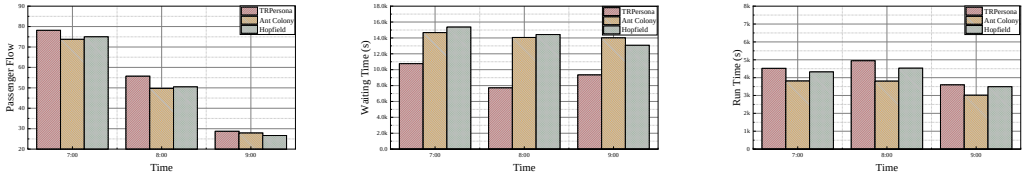
1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics



(a) Passenger flow (b) Waiting time (c) Travel time

Fig. 6: Comparison of TRProfiling and real route.



(a) Passenger flow (b) Waiting time (c) Travel time

Fig. 8: Comparison of TRProfiling and other methods.

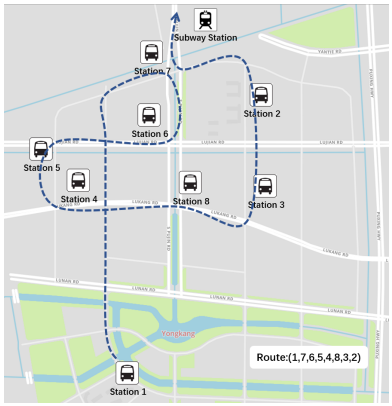


Fig. 5: The application area of the experiments.

TABLE III: Description of APP order data.

Field	Annotation
OrderID	ID of each order
Type	Orders type
RegionID	Region ID
PassengerID	ID of each passenger
CreateDate	Order creation date and time
CheckTicketDate	Passenger boarding date and time
UpStopID	Station ID of passengers boarding
RideCount	the number of passengers

TABLE IV: Description of shared bus GPS data.

Field	Annotation
BusID	Shared bus ID
Time	Record time
Timestamp	Record timestamp
Latitude	The latitude of shared bus
Longitude	The Longitude of shared bus
RegionID	Region ID
StationID	ID of the last station
LastLatitude	Last latitude of shard bus
LastLongitude	Last longitude of shared bus
SeatNum	The number of seats of shared bus

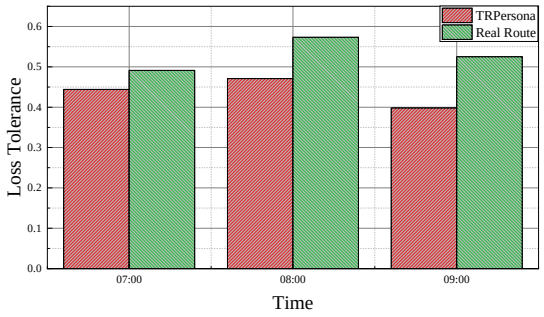


Fig. 7: Comparison of Loss Tolerance.

promoting the efficiency of going to work, which is more urgent than going off work. Furthermore, we filter out the data of missing values, payment in cash and canceled orders. Based on the operating rules that it takes the shared buses nearly 20 minutes for a trip, the period is divided into small time intervals of 10 minutes from 6:40 am to 9:40 am. From the GPS trajectory data, the travel time and distance between each station can be calculated. In our paper, we use station i and j as the origin station and the destination respectively, and the OD pair (Station i, Station j) is determined.

C. Optimal Route Evaluation

We conduct a series of simulation experiments and compare with the actual routes with four metrics, i.e., passenger flow, travel time, waiting time, and loss tolerance, to evaluate the practicability of our method.

We choose several rush hours in the morning and predict the results shown in Fig. 6. To sum up, the travel time of our method is consistent with actual route data. However, the value of passenger flow and passenger waiting time are much better

1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

Algorithm 1 Multi-Constraints Evolution Algorithm

Input: $G(V, E)$,The graph of the stations with a given OD pair
TDM: Travel Distance Matrix
PFM: Passenger Flow Matrix
TTM: Travel Time Matrix
WTM: Waiting Time Matrix
SN: The number of seats
Output: R: The Optimal Route

P_{num} = 0, P_{num} is the number of passengers that has already got on the bus
Select R_{init}, R_{init} is the initial route selected from the actual routes that satisfy three constraints (7), (8) and (9)
Calculate R_{init_Z} based on Equation. (4) and Equation. (5)
while Algorithm is not convergent **do**
  while $P_{num} \leq SN$ AND $G(V, E)$!= $\emptyset$ **do**
    for each station $\in$ candidate stations **do**
      Calculate P with Equation. (19)
    end for
    select station s_i based on probability P
    add s_i to R
    update P_{num}
    remove s_i from $G(V, E)$
  end while
  Calculate three constraints of R: R_{L_t}, R_{D_t}, R_{S_r} of Equation. (9), Equation. (10), Equation. (11)
  if $R_{L_t} < m$ AND $R_{D_t} < n$ AND $k < R_{S_r} \leq 1$ **then**
    Calculate R_Z based on Equation. (4) and Equation. (5)
    if R_Z < R_{init_Z} **then**
      $R_{init} \leftarrow R$
      $R_{init_Z} \leftarrow R_Z$
      update $P_{SUR_{flow}}$,$P_{SUR_{seat}}$,$P_{T_{time}}$,$P_{W_{time}}$ based on the Equation. (15), Equation. (16), Equation. (17), Equation. (18)
    end if
  end if
end while
return R

than real routes. Furthermore, our optimized method reduces the passenger waiting time by nearly a half on some departure time spots. It illustrates that the comprehensive performance of TRProfiling has a large advantage compared with the actual routes.

Fig. 7 shows the comparison of loss tolerance between the optimal route and real route. It is quite obvious that the loss probability of passengers significantly reduces after optimization. In other words, the optimized route strategy not only greatly improve the revenues of bus companies, but also it is suitable for satisfying different passengers' needs.

D. Comparison of Route Planning Approach

To further verify the effectiveness of our proposed TRProfiling, we conduct a comparison among different schemes, i.e., ant colony algorithm [32], [33], and Hopfield neural network [34].The ant colony algorithm is one of the most classical bionics algorithms proved to be efficient in many fields. The Hopfield neural network is an artificial intelligence algorithm which is applied in solving combinatorial optimization problems in recent years [22]. As shown in Fig. 8, TRProfiling is obviously better than other methods in reducing the passenger waiting time. Besides, TRProfiling also has advantages to improve passenger flow. Ant colony method, as the metaheuristic approximation method, allows parallel implementation due to its inherent nature [35], which is superior to TRProfiling in reducing travel time. However, it is acceptable for TPProfiling considering the improvement in other aspects. Therefore, the performance of TRProfiling is better than the other algorithms in the dynamic route planning for shared buses in the aspect of meeting passenger demands. TRProfiling can accurately characterize the travel demands of passengers and then use the MCEA algorithm to select the optimal routes considering these demands.

V. Conclusion

To overcome the data scarcity issue and optimize shared bus services, we propose a heterogeneous mobile crowdsourced data-based shared bus profiling scheme, TRProfiling. First, we design an MCS-based data generation and collection solution and cooperate with an Internet company, Yitong Innovation Techonology (Dalian) Co., to implement a mobile APP, Futurefleet, to collect shared bus data. Two datasets, order data and GPS, covering data for more than six months, are acquired. Then based on the analysis of these datasets, we construct TP to describe resident travel requirements and adopt machine learning algorithms to instantiate them. Following that, MCEA is presented to select the optimal route for shared buses, considering multi-cost, i.e., the multiple features of TP. At last, a series of experiments are performed to demonstrate the effectiveness of our method. It also indicates that TRProfiling has immense value in industrial applications and provides construction suggestions to the development of shared buses.

In order to make the TRProfiling more effective, we will devote into considering more costs such as air pollution and traffic jams to enrich TP in further work. We will also focus on optimizing the algorithm to improve the efficiency of TRProfiling scheme.

VI. Acknowledgments

The authors would like to thank Zhenhuan Fu for his help with the experiments.

References

[1] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial internet of things: Challenges, opportunities, and directions," *IEEE Transactions on Industrial Informatics*, vol. PP, no. 99, pp. 1–1, 2018.
[2] I. Akkaya, P. Derler, S. Emoto, and E. A. Lee, "Systems engineering for industrial cyber–physical systems using aspects," *Proceedings of the IEEE*, vol. 104, no. 5, pp. 997–1012, 2016.
[3] M. R. Pedersen, L. Nalpantidis, R. S. Andersen, C. Schou, S. Bøgh, V. Krüger, and O. Madsen, "Robot skills for manufacturing: From concept to industrial deployment," *Robotics and Computer-Integrated Manufacturing*, vol. 37, pp. 282–291, 2016.

1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TII.2019.2947063, IEEE Transactions on Industrial Informatics

[4] C. Perera, C. H. Liu, S. Jayawardena, and M. Chen, "A survey on internet of things from industrial market perspective," *IEEE Access*, vol. 2, pp. 1660–1679, 2014.

[5] L. Liu, "Services computing: from cloud services, mobile services to internet of services," *IEEE Transactions on Services Computing*, no. 5, pp. 661–663, 2016.

[6] A. Gupta, W. Thies, E. Cutrell, and R. Balakrishnan, "mclerk: enabling mobile crowdsourcing in developing regions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2012, pp. 1843–1852.

[7] H. Chen, B. Guo, Z. Yu, L. Chen, and X. Ma, "A generic framework for constraint-driven data selection in mobile crowd photographing," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 284–296, 2017.

[8] S. Foell, G. Kortuem, R. Rawassizadeh, M. Handte, U. Iqbal, and P. Marron, "Micro-navigation for urban bus passengers: using the internet of things to improve the public transport experience," in *Proceedings of the First International Conference on IoT in Urban Space*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2014, pp. 1–6.

[9] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *nature*, vol. 453, no. 7196, p. 779, 2008.

[10] H. Ma, D. Zhao, and P. Yuan, "Opportunities in mobile crowd sensing," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 29–35, 2014.

[11] D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *Computer*, vol. 44, no. 7, pp. 21–28, 2011.

[12] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Y. Ma, "Recommending friends and locations based on individual location history," *Acm Transactions on the Web*, vol. 5, no. 1, p. 5, 2011.

[13] M. Ye, P. Yin, W. C. Lee, and D. L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *International Acm Sigir Conference on Research & Development in Information Retrieval*, 2011, pp. 325–334.

[14] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone:an end-to-end participatory urban noise mapping system," in *ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2010, pp. 105–116.

[15] J. Chen, C. Wang, and J. Wang, "Will you" reconsume" the near past? fast prediction on short-term reconsumption behaviors," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[16] D. Rafailidis and A. Nanopoulos, "Repeat consumption recommendation based on users preference dynamics and side information," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 99–100.

[17] J. Goldman, "Participatory sensing : A citizen-powered approach to illuminating the patterns that shape our world," *Foresight & Governance Project White Paper*, pp. 117–134, 2009.

[18] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.

[19] O. Wolfson, Y. Zheng, and S. Ma, "T-share: A large-scale dynamic taxi ridesharing service," in *IEEE International Conference on Data Engineering*, 2013, pp. 410–421.

[20] M. Li, L. Chen, G. Cong, Y. Gu, and G. Yu, "Efficient processing of location-aware group preference queries," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 559–568.

[21] Y. Zhou, Q. Luo, H. Chen, A. He, and J. Wu, "A discrete invasive weed optimization algorithm for solving traveling salesman problem," *Neurocomputing*, vol. 151, pp. 1227–1236, 2015.

[22] L. García, P. M. Talaván, and J. Yáñez, "Improving the hopfield model performance when applied to the traveling salesman problem," *Soft Computing*, vol. 21, no. 14, pp. 3891–3905, 2017.

[23] R. Lahyani, M. Khemakhem, and F. Semet, "A unified matheuristic for solving multi-constrained traveling salesman problems with profits," *EURO Journal on Computational Optimization*, vol. 5, no. 3, pp. 393–422, 2017.

[24] T. Liebig, N. Piatkowski, C. Bockermann, and K. Morik, "Dynamic route planning with real-time traffic predictions," *Information Systems*, vol. 64, pp. 258–265, 2017.

[25] S. Wang, W. Lin, Y. Yang, X. Xiao, and S. Zhou, "Efficient route planning on public transportation networks: A labelling approach," in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. ACM, 2015, pp. 967–982.

[26] J. Dibbelt, T. Pajor, and D. Wagner, "User-constrained multimodal route planning," *Journal of Experimental Algorithmics (JEA)*, vol. 19, pp. 3–2, 2015.

[27] S. Shang, L. Chen, Z. Wei, C. S. Jensen, K. Zheng, and P. Kalnis, "Parallel trajectory similarity joins in spatial networks," *The VLDB JournalâĂŤThe International Journal on Very Large Data Bases*, vol. 27, no. 3, pp. 395–420, 2018.

[28] C. Chen, D. Zhang, N. Li, and Z.-H. Zhou, "B-planner: Planning bidirectional night bus routes using large-scale taxi gps traces," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 4, pp. 1451–1465, 2014.

[29] B. Yang, C. Guo, C. S. Jensen, M. Kaul, and S. Shang, "Stochastic skyline route planning under time-varying uncertainty," in *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*. IEEE, 2014, pp. 136–147.

[30] P. Zhou, Y. Zheng, and M. Li, "How long to wait? predicting bus arrival time with mobile phone based participatory sensing," *IEEE Transactions on Mobile Computing*, vol. 13, no. 6, pp. 1228–1241, 2014.

[31] C. M. Grinstead and J. L. Snell, *Introduction to probability*. American Mathematical Soc., 2012.

[32] M. M. Hamed, H. R. Al-Masaeid, and Z. M. B. Said, "Short-term prediction of traffic volume in urban arterials," *Journal of Transportation Engineering*, vol. 121, no. 3, pp. 249–254, 1995.

[33] M. Dorigo and M. Birattari, "Ant colony optimization," in *Encyclopedia of machine learning*. Springer, 2011, pp. 36–39.

[34] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[35] R. S. Parpinelli, H. S. Lopes, and A. A. Freitas, "Data mining with an ant colony optimization algorithm," *IEEE transactions on evolutionary computation*, vol. 6, no. 4, pp. 321–332, 2002.

1551-3203 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.