CommuteShare: A Ridesharing Service for Daily Commuters Using Cross-domain Urban Big Data

Xiaoliang Fan[1,6], Chang Xu[2], Fang Tang[3], Jianzhong Qi[4], Xiao Liu[5], Longbiao Chen[1,6], Cheng Wang[1,6,*]
[1] Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, Xiamen, China
[2] Software School, Xiamen University, Xiamen, China
[3] Xiamen Branch, Industrial Bank Co., LTD, Xiamen, China
[4] School of Computing and Information Systems, University of Melbourne, Melbourne, Australia
[5] School of Information Technology, Deakin University, Melbourne, Australia
[6] Digital Fujian Institute of Urban Traffic Big Data Research, Xiamen University, Xiamen, China
Email: xfb_fxl@xm.gov.cn, 528994772@qq.com, 052310@cib.com.cn, jianzhong.qi@unimelb.edu.au, xiao.liu@deakin.edu.au, {longbiaochen, cwang}@xmu.edu.cn

***Abstract*—Existing ridesharing services have focused on on-demand trip matching, which resembles traditional taxi dispatching. This may encourage more private vehicles on the road, which aggravate traffic congestions in peak hours rather than alleviating them. We propose *CommuteShare*, a novel ridesharing service for daily commuters that encourages long-term ridesharing among commuters with similar commuting patterns, to increase the traffic efficiency in peak hours. We first identify *commuting private vehicles* (CPVs) from traffic records and model their commuting patterns. We then design a dynamic model to formulate the intention level of a CPV driver to offer a ride based on the spatio-temporal convenience and dynamic traffic conditions. Based on the commuting patterns of the CPVs and the dynamic model of the CPV drivers, we propose a ridesharing algorithm to compute ridesharing matches among CPVs. We perform extensive experiments on three real-world cross-domain urban big datasets from a major city of China. Experimental results show that, using the proposed CommuteShare service, over 5,300 private vehicles can be reduced daily on average during morning peak hours, with a reduction of 7-minute average waiting time for the riders.**

***Keywords*—ridesharing, services computing, urban computing.**

I. Introduction

Heterogeneous urban big data contributed by both the crowd and the large number of sensors brings opportunities to create innovative services to solve a variety of urban problems via understanding individuals' behaviors and their collective patterns. In this study, we are interested in alleviating traffic congestions using urban big data. For a long time, public transportation systems (e.g., buses and subways) have been a main way to alleviate traffic congestions. However, there is still a common use of private vehicles for commuting, due to the long waiting time and far walking distances of public transportation systems in peak hours. A survey shows that *commuting private vehicles* (CPVs) has an average occupancy rate (i.e., the number of travelers per trip) of 1.17. Such a low occupancy rate of CPVs has been considered as a major source of traffic inefficiency in urban transportation [1].

Ridesharing (a.k.a. carpooling) services are designed to overcome the traffic inefficiency of non-shared rides by offering the vacant seats to additional passengers [2]. Many ridesharing service providers, e.g., Uber[1] and Didi Chuxing[2], have mobile APPs to match drivers' offers with riders' requests. Traditionally, there are two main types of ridesharing services [2]: (1) *on-demand ridesharing*, which is for one-time trips and requires a real-time response without considering any historical mobility patterns of the riders; and (2) *tailored ridesharing*, which is for a group of passengers with similar and repeating mobility patterns. On one hand, on-demand ridesharing services rely much on a large number of CPVs on the road, which may aggravate the traffic congestion rather than alleviating it. On the other hand, tailored ridesharing services, such as airport shuttles, often operate with a fixed route and a fixed schedule. Thus, they may have a long detour distance to pick up all the riders, which causes a low user satisfactory.

Inspired by studies on human mobility with large-scale datasets such as GPS [4], Call Detail Records [5] and check-ins [6], we utilize a large-scale *vehicle license plate recognition* (VLPR) dataset [7] to follow a tailored ridesharing scheme by reducing the number of vehicles on the road while incorporating the flexibility of on-demand ridesharing. We propose *CommuteShare,* an improved ridesharing service for daily commuters that encourages stable ridesharing companions. First, we identify CPVs from massive traffic records and model their commuting patterns. Second, a ridesharing algorithm is proposed to compute ridesharing matches among CPVs. An extensive empirical study is conducted on three real-world cross-domain urban datasets from Xiamen, a major city of China. The datasets contain 4 million commuting trips in May 2016, 134,721 traffic accidents during January and September 2016, and hourly weather data in May 2016. The experimental results show that, using the CommuteShare system, over 5,300 private vehicles can be reduced on daily average during 7am

* Correspondence author.
[1] Uber, https://www.uber.com
[2] Didi Chuxing, http://www.xiaojukeji.com

978-1-5386-7247-1/18/$31.00 ©2018 IEEE
DOI 10.1109/ICWS.2018.00046

and 9am in the morning peak hours[3], with only a 7-minute average waiting time for the riders.

The rest of this paper is organized as follows. Section II presents related studies. Section III describes the cross-domain urban datasets and observations. Section IV details the proposed algorithms. Section V shows the experimental results. Section VI concludes the paper.

II. Related Work

Ridesharing is a process of at least two travelers sharing a ride in a vehicle with respect to regular itineraries. The benefits of ridesharing services are manifold [3], including waiting time saving, travel cost reduction, etc. Two types of ridesharing are [2]: on-demand ridesharing, and tailored ridesharing.

Wang et al. [7] propose to incorporate the human mobility mechanism into unlicensed taxis detection from massive citywide vehicles.

Tailored ridesharing [6], such as tailored buses, is designed for travel companions with similar mobility patterns (e.g., travelling on the same route). Tailored ridesharing has a high occupancy rate and is environmental friendly. However, it is inflexible and forces passengers to comply with fix routes and schedules of the tailored services. To avoid those limitations above, we aim to design a ridesharing service that balances the user satisfaction (i.e., a high ridesharing match and a short waiting time) and the social benefit (i.e., a large number of vehicles reduced from the road).

III. Datasets and Observations

We first describe the data used for our study.

A. Cross-domain Urban Big Datasets

We use datasets from the following three domains that impact the traffic efficiency.

(1) Vehicle License Plate Recognition (VLPR) Data

VLPR devices can recognize a vehicle's license plate number attributing to advanced techniques in image processing and pattern recognition [7]. VLPR devices deployed on city-wide road networks (i.e., 439 devices in Xiamen) are capable of generating a large-scale mobility dataset, reflecting the city pulse of traffic congestions, accidents, as well as vehicle moving patterns. For example,

We obtained two VLPR datasets in Xiamen: the *VLPR records* dataset and the *VLPR devices* dataset. The VLPR records dataset records when and where a vehicle passes a VLPR device, while VLPR devices dataset records the geographic information (i.e., longitude and latitude) and types of VLPR devices.

(2) Multi-sourced Traffic Data

We consider two types of traffic data that may impact the ridesharing efficiency: traffic accident and traffic condition. *Traffic accidents* are considered as one of the major causes of traffic congestions, and dataset contains 134,721 traffic accidents during the first nine months of 2016 in Xiamen. Each accident record has four fields: accident ID, accident time, accident coordinates, and crash type (i.e., single vehicle, side-wipe, or rear-end). *Traffic condition* represents the level of traffic congestions, and our dataset contains the average travel time for CPVs in morning peak hours. Fig. 1 shows, CPVs spend more travel times on Monday than other working days, indicating that Monday mornings are more congested.

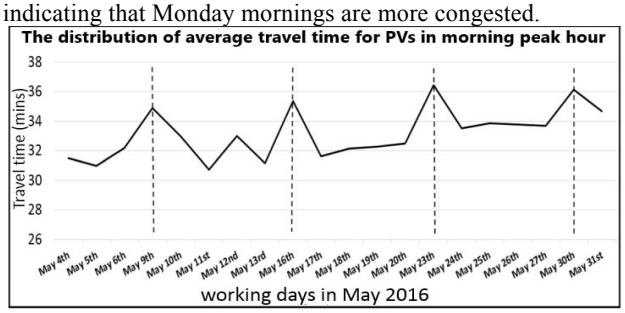

Figure 1. Data observation of traffic condition.

(3) Weather Forecasting Data

Weather conditions have a significant impact on traffic congestion and anomalies [8]. Precipitation, such as the amount of rainfall/snowfall and the duration, is the most important weather factor to traffic congestion. Other weather conditions such as fogs can reduce visibility. We collected a weather dataset that consists of historical hourly weather data[4] in May 2016 in Xiamen, including attributes such as precipitation, visibility, temperature, atmospheric pressure, etc.

B. Data Pre-processing

First, private vehicles are recognized from the VLPR dataset, according to both the color of the license plate. Second, we extract key attributes from the VLPR dataset, which include the following 10 attributes: license plate number, license plate color, lane number, device id, device type, device direction, passed time, device location, longitude, and latitude. Third, we filter out redundant and incomplete VLPR records, e.g., a VLPR device may take redundant pictures of the same vehicle during the congestion. Finally, trajectories of each private vehicle are generated sequentially.

C. Observations

We make the following observations on the daily mobility patterns of *commuting private vehicles* (CPVs).

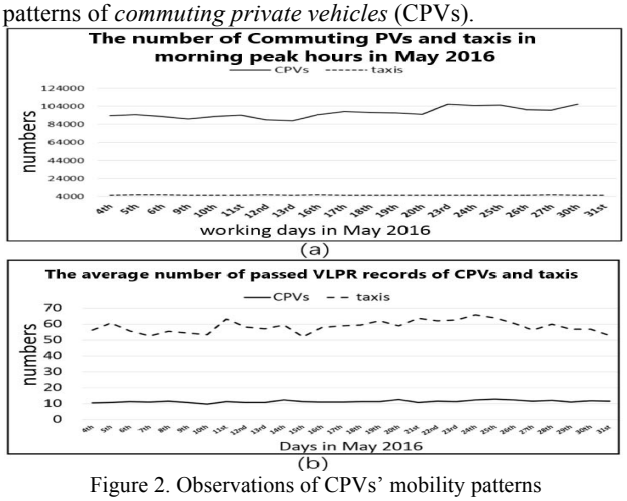

Figure 2. Observations of CPVs' mobility patterns

[3] We focus on identifying ridesharing companions for morning peak hours only and leave the evening commuting problem for the future work.

[4] Open weather data, http://rp5.ru/Weather_in_the_world

(a) The number of CPVs is huge. We compared the numbers of CPVs and taxis in Fig. 2(a), and the number of CPVs in working days is much larger than that of taxis.

(b) The travel distance of CPVs is short. In Fig. 2(b), we find that a CPV has a very short travel distance (around 10 VLPR records in each trip) in morning peak hours.

These observations imply the opportunity to create a ridesharing service for CPVs users to reduce the number of non-shared CPVs on the road in peak hours.

IV. The CommuteShare Method

A. Proposed Framework

We propose a framework to compute the optimal matching of CPVs with riders leveraging the spatio-temporal closeness of the commuting trips in Fig. 3.

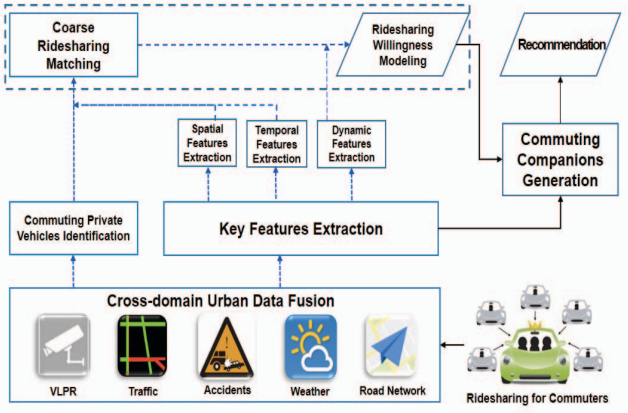

Figure 3. Overview of the proposed framework

B. Ridesharing Algorithm

We detail the proposed CommuteShare algorithm below. We use all workdays in a month (i.e., 20 workdays in May 2016) as the range of observations.

Stage 1: CPV Identification

First, by considering the periodicity and repeatability of commuting patterns of CPVs, we exclude inactive private vehicles which have VLPR records in less than 15 working days per month.

Second, in Fig. 4, we infer potential home and work locations (VLPR device locations close to home or work place) according to the staying duration between two VLPR records. Intuitively, a CPV stays the longest duration at home overnight, and spends the longest duration at work in the daytime.

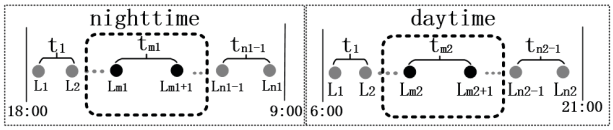

Figure 4. Identify diurnal stay location and overnight stay location

Third, we infer more accurately the home and work locations from the potential locations, as there may be more than one home or work locations identified by the heuristics used in the second step.

Finally, from 1,204,790 private vehicles, we identify 18,944 CPVs, and the trajectories between two closest VLPR devices to their home/work locations are considered as commuting trajectories.

Stage 2: Feature Extraction

Feature extraction includes the following three steps.

(1) *Spatial feature extraction*. For each CPV, we extract its home and work coordinates as the spatial feature.

(2) *Temporal feature extraction*. We use the departure time and arrival time during the observation period (i.e., 20 workdays in our experiment) as the temporal features.

(3) *Dynamic feature extraction*. We use traffic condition, accidents, and weather in each workday as dynamic features.

In summary, we extract 124 features, including the locations of home or work (four coordinates), departure and arrival times in 20 working days (40 time values), Monday-or-not, number of accidents, visibility, and precipitation in 20 days (80 values).

Stage 3: Dynamic Ridesharing Matching

We perform matching in two steps.

(1) *Coarse ridesharing matching*. We first consider the temporal features and spatial features of CPVs which form feature vectors of 44 dimensions. We apply the k-means algorithm on these feature vectors to cluster the CPVs to find the ones sharing similar commuting patterns.

(2) *Ridesharing based on willingness*. In order to identify pick-up/drop-off coordinates and the departure/arrival times for rideshared CPVs in a cluster, we need to generate the center vector for each CPV cluster, which is computed as the average of all CPVs in a cluster. For example, we denote all feature vectors in a cluster as $FV_1, FV_2, \dots, FV_n$, and each feature vector can be denoted as $(FV_{i1}, FV_{i2}, \dots FV_{im})$, where $i = 1, 2, \dots, n$ and m is the number of dimensions of feature vectors. The center vectors of the cluster is computed as $CV = (CV_1, CV_2, \dots, CV_m)$, $CV_i = \frac{\sum_{f=1}^{n} FV_{fi}}{n}$. As a result, the center vector includes the information of pick-up/drop-off coordinates and departure/arrival time in each working day for rideshared CPVs in a cluster. The pseudo code of the above procedure is as follows.

Algorithm. Ridesharing algorithm

Input: $seats$, number of ridesharing seats

c, inputted cluster (single cluster)

k, number of resultant clusters, $k = size(c)/seats$

Output: $\boldsymbol{C}$, resultant clusters

1) use AGNES algorithm to generate an initial cluster result $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \dots, \boldsymbol{c}_k\}$;
2) ***for*** each cluster $\boldsymbol{c}_i \in \boldsymbol{C}$ ***do***
3) ***if*** size($\boldsymbol{c}_i$) > seats ***do***
4) k ← size($\boldsymbol{c}_i$)/seats;
5) ***go to*** 1);
6) ***else***
7) save $\boldsymbol{c}_i$;
8) ***end if***
9) ***end for***

V. Performance Evaluation

A. Experimental Settings

The experiments are run with MATLAB (R2014a) on an ASUS K55V computer with 64-bits Windows 7 system, 32 GB RAM and an Intel Core I7 3.6 GHz CPU.

The performance of CommuteShare is measured by three metrics: 1) average waiting time of riders; 2) average rideshare accuracy; and 3) number of vehicles reduced.

B. Baselines

- ***K-means based on Squared Euclidean Distance*****,** the distance measurement function is denoted as: $d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|^2$.
- ***K-means based on BlockCity Distance***, the distance measurement function is denoted as: $d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$.
- ***K-means*****++**, i.e., an improvement of k-means, which uses initial centers of clusters as far from each other as possible to achieve a better clustering result.
- ***AGNES*****,** i.e., a classic condensed hierarchical clustering algorithm, which uses each target as a cluster at the beginning, and then merges the clusters progressively.

C. Evaluation and Results

Two evaluation metrics (i.e., average waiting time of riders and average rideshare accuracy) are presented to evaluate the performance of CommuteShare, compared with the baseline methods.

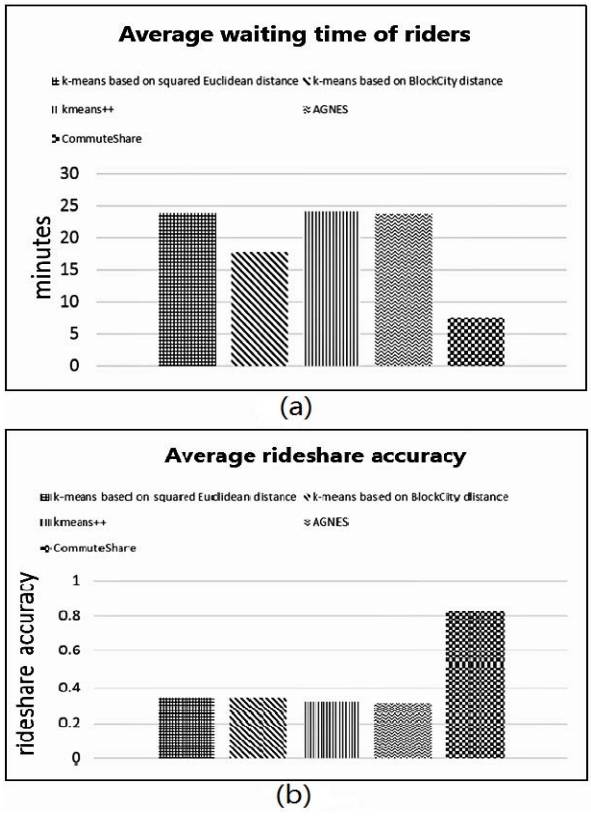

Figure 5. The comparison of (a) average waiting time of riders, and (b) average rideshare accuracy.

Fig. 5(a) shows that the average waiting time of CommuteShare is 7 minutes, which is significantly less than baseline methods. Fig. 5(b) indicates that the average rideshare accuracy of CommuteShare is significantly higher than baselines.

VI. Conclusion

We presented a novel ridesharing service, CommuteShare, which is able to match rideshare companions with a high accuracy for daily commuters. Specifically, CommuteShare significantly outperforms the baseline methods in improving both riders' satisfaction and social benefits: 1) CommuteShare achieves higher rideshare accuracy (an average daily rideshare accuracy of 83%) than the baseline methods; 2) CommuteShare can reduce over 30% of commuting private vehicles during morning peak hours of workdays.

Future works may include: 1) extending the algorithm to enable dynamic ridesharing in evening peak hours where there are many complicated alternatives for the commuters rather than going home directly (e.g., going shopping, dining out, visiting a bar, etc.); and 2) employing deep neural networks to uncover the complexity in ridesharing.

Acknowledgement

The work is supported by grants from National Natural Science Foundation of China (61300232), China Postdoc Foundation (2015M580564), and Australian Research Council (Discovery Project DP180103332).

References

[1] D. M. Kammen and D. A. Sunter, "City-integrated renewable energy for urban sustainability," Science, vol. 352, May. 2016, pp. 922–928, doi:10.1126/science.aad9302.

[2] M. Furuhata, M. Dessouky, F. Ordóñez, M.-E. Brunet, X. Wang, and S. Koenig, "Ridesharing: The state-of-the-art and future directions," Transportation Research Part B: Methodological, vol. 57, Nov. 2013, pp. 28–46, doi:10.1016/j.trb.2013.08.012.

[3] N. D. Chan and S. A. Shaheen, "Ridesharing in North America: Past, Present, and Future," Transport Reviews, vol. 32, Jan. 2012, pp. 93–112, doi:10.1080/01441647.2011.621557.

[4] D. Zhang, L. Sun, B. Li, C. Chen, G. Pan, S. Li, and Z. Wu, "Understanding Taxi Service Strategies from Taxi GPS Traces," IEEE Transactions on Intelligent Transportation Systems, vol. 16, Feb. 2015, pp. 123–135, doi:10.1109/tits.2014.2328231.

[5] B. Cici, A. Markopoulou, E. Frias-Martinez, and N. Laoutaris, "Assessing the potential of ride-sharing using mobile and social data," in Proc. ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 14), 2014, pp. 201-211, doi:10.1145/2632048.2632055.

[6] Y. Han, G. Wang, J. Yu, C. Liu, Z. Zhang, and M. Zhu, "A Service-Based Approach to Traffic Sensor Data Integration and Analysis to Support Community-Wide Green Commute in China," IEEE Transactions on Intelligent Transportation Systems, vol. 17, Sep. 2016, pp. 2648–2657, doi:10.1109/tits.2015.2498178.

[7] S. Du, M. Ibrahim, M. Shehata, and W. Badawy, "Automatic License Plate Recognition (ALPR): A State-of-the-Art Review," IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, Feb. 2013, pp. 311–325, doi:10.1109/tcsvt.2012.2203741.

[8] M. J. Koetse and P. Rietveld, "The impact of climate change and weather on transport: An overview of empirical findings," Transportation Research Part D: Transport and Environment, vol. 14, May 2009, pp. 205–221, doi:10.1016/j.trd.2008.12.004.