Trajectory Topic Modelling to Characterize Driving Behaviors with GPS-Based Trajectory Data

Lyuchao Liao[1,2], Jianping Wu[1], Fumin Zou[2], Jengshyang Pan[3], Tingting Li[1]

[1] Department of Civil Engineering, Tsinghua University, China
[2] Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, China
[3] Fujian Provincial Key Lab of Big Data Mining and Applications, Fujian University of Technology, China
fjachao@gmail.com, jianpingwu@tsinghua.edu.cn, {fmzou, jengshyangpan}@fjut.edu.cn, ttlithu@163.com

Abstract

The rapid accumulation of large-scale driving data represents an opportunity to improve our understanding of driving behavior patterns and driver traveling intentions. However, limited efforts have been devoted to understanding these patterns and the travel intentions behind them. This study proposes a new trajectory topic model (TTM) to explore latent driving patterns from driving trajectory data and to qualitatively analyze drivers' main traveling intentions. These trajectory data were collected from more than 150,000 commercial vehicles in Fujian Province, China. After data preprocessing, the TTM was then established to decompose trajectory data into various topics with corresponding probabilities, which were correlated to drivers' preferences. Several experiments conducted in Fuzhou City were performed to evaluate the feasibility and efficiency of the TTM using a real trajectory dataset. The results show that the TTM could effectively mine users' driving behavior patterns with topic probability. The model would enable us to understand the context in which drivers travel and learn their individual preferences. It is also beneficial in that it can predict drivers' behaviors, analyze traffic patterns in an entire city, and even help autonomous vehicles to learn from drivers.

Keywords: Trajectory data mining, Trajectory topic model, Latent Dirichlet Allocation (LDA), Driving behavior analysis, Traveling intention

1 Introduction

Trip destinations and the routes are, in the context of driving behaviors, key to enhancing safety and convenience for drivers. By obtaining a trip destination, a driver assistance system can automatically derive an optimized travel route [1-3] and provide route-specific traffic information [4] or good advice to drivers [5]. In addition, with the trip route known in advance, it is possible, as is shown by researchers from Nissan, to improve hybrid fuel economy by up to 7.8% [6]. Furthermore, to predict destinations and then optimize routes is also beneficial in predicting short-term traffic [7-10], conducting transportation planning [11], and advertising commercial products by targeting customers along certain routes [12]. For example, given the trip destination, advertisements can be targeted at customers who are likely to be along the route to that destination.

The potential advantages of analyzing driving behavior have spurred considerable research in recent years. Existing studies can mainly be divided into two research categories: geometric probability inference (GPI) and semantic behavior analysis (SBA). By using GPI-based approaches, users' positions are projected onto segments of a road network. Then their driving origin-destination (OD) is predicted by performing a probability transition of the road-segmentation sequences [13-16]. By using SBA-based approaches, driving OD behaviors are explored by clustering users' patterns on the semantic level [17-19]. To date, GPI-based methods have mainly focused the low-level features of spatiotemporal density instead of higher-level semantic patterns to understand the traveling intentions of drivers. SBA-based methods, which focus mainly on micro-behavior analysis, have not investigated the route behaviors, and thus, predicting routes becomes difficult in the long term. A review of GPI and SBA are presented in Section2: Related Studies. The Latent Dirich-let Allocation (LDA) model [20] attempts to understand how a document is generated by assuming that a document is a mixture of different topics and that each word is generated in relation to one of these topics. By this analogy, driving trajectory can be regarded as a mixture of topics that are related to different traveling intentions. These topics are produced with certain probabilities of the distribution of vehicles on spatial grids. If the drivers' intrinsic behaviors are identified, this topic pattern analysis on trajectory data can be used to produce robust traveling patterns from historical events. However, it is still challenging to learn driving patterns

*Corresponding Author: Lyuchao Liao; E-mail: fjachao@gmail.com
DOI: 10.3966/160792642018051903017

from massive and noisy historical trajectory data of traveling events.

Topic pattern analysis of driving behaviors adopts this idea from the semantic mining of massive texts. The idea is to model all the term data in a fixed order and to explore the high-level patterns of driving behavior with semantic correlations. Although the text data are out of order, the topic pattern model makes it possible to get the potential semantic information covering various text-equivalents, such as abbreviation, equivalent expression, and synonyms. Similarly, the topic pattern model can also be employed to explore the semantic similarities between different geospatial regions of interest (ROIs) from massive trajectory data. For example, if the passing location of a car is known, the probability of it passing another car could be estimated with semantic similarity.

In this paper, as an unsupervised model, a trajectory topic model (TTM) is proposed to establish high-level behavior topics of traveling events from massive trajectory data. As an expression of traveling events, trajectory data exert a powerful function to reveal drivers' traveling intentions and their driving preferences. The TTM is thus used to quantitatively analyze drivers' traveling intentions and to predict travel routes or trip destinations. The contribution of this paper is three-fold. First, this paper proposes a novel method to transform and decompose massive and noisy trajectory data into uniform vectors for the easy application of a large-scale data analysis method. Second, it presents behavior topics from driving trajectory data that are related to specific traveling intentions. Third, this paper proposes a new TTM to semantically analyze driving behaviors from massive and noisy trajectory data in order to understand and predict behavior patterns of traveling events.

2 Related Studies

The existing research analyzing driving behaviors can be classified into GPI and SBA.

2.1 GPI

GPI is a method to determine the probability that a vehicle will travel along a given segment of road network and then to infer trip destinations with maximum probability. After reconstructing vehicle trajectories using GPS data, some probability-learning models, such as the probabilistic filter model [21], path inference filter (PIF) [22], and frequent trajectory patterns (FTP) [15], have been introduced to estimate the behavior of traffic participants and to anticipate their future paths [13-14]. Having clustered staying locations (positions at which a vehicle is at rest) in the driving paths and matching these locations to geometric POIs, researchers extracted drivers' activities from historical trajectories and identified the purposes of their trips [23]. Furthermore, researchers transformed paths composed of sparse GPS data to edges of a road network, expecting to learn users' path preferences and to predict specific routes [16, 24-25]. With map-matching in advance, a tensor-formed model is also employed to represent drivers' spatiotemporal features and to capture dynamic crowd patterns [26-29]. Furthermore, some classical algorithms, such as dynamic Bayesian network (DBN) [30], hidden Markov models (HMMs) [31-32], neural networks, [33] and evolution models [33-35], have captured drivers' behaviors from trajectory data. These methods can analyze users' preferences, but they focus more on the low-level features of spatiotemporal density and thus neglect high-level semantic patterns.

2.2 SBA

Only a few studies have focused on the semantic patterns associated with predicting driver trajectories and other related data. Given the importance of behavior intention analysis, some topic models are proposed to capture users' movements and their patterns from check-in data [36-37] and other society data, such as online forums, blogs, and Weibo (Twitter-like websites in China) [38]. Virtually, data collected on actual vehicles including speed, acceleration, etc. can reveal drivers' latent behavior patterns and predict their destinations [17]. With the aim of micro-behavior analysis, segment algorithms are proposed to assign trip sections to high-level semantic behaviors, such as braking, turning, accelerating, coasting, and standing still [39]. In the same way, trajectories are also divided into sections to find anomalous movements and classify levels of danger [19]. Moreover, methods based on this topic model were employed to exploit users' periodic behaviors in analogy to co-occurrence analysis in text mining [20, 40]. These existing works on micro-behavior analysis, however, have not taken path behavior into account, so they are unable to predict paths in long term [41].

3 Description of Trajectory Data

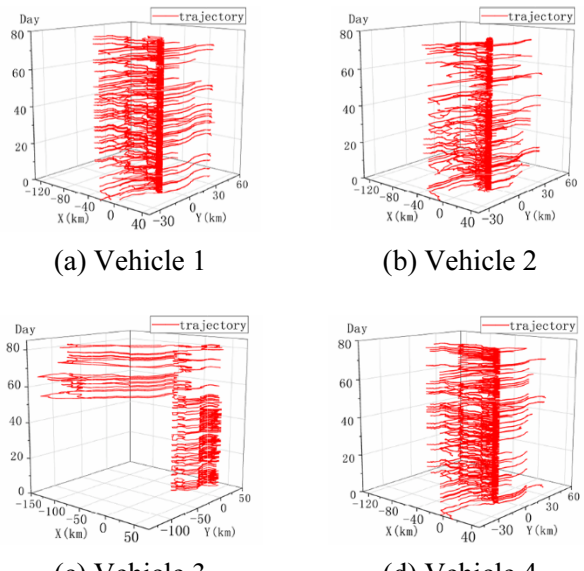
Traffic trajectory data are sequences of spatiotemporal locations tagged with time stamps that record the movement and status of vehicles. The trajectory dataset used in this paper was collected from over 150,000 vehicles by the Fujian Provincial Department of Transportation in China. It includes more than 10 types of commercial vehicle (i.e., taxis, heavy trucks, urban buses, intercounty buses, intercity buses, interprovince buses, travel buses, semi-trailer trucks, vehicles transporting dangerous goods, and others). The trajectory data mainly consist of a vehicle ID, a time stamp, the vehicle's current location from GPS (longitude and latitude), speed, vehicle direction.

The sampling frequency ranges from 5 to 60 seconds depending on the type of vehicle. The recorded trajectory data are demonstrated in Figure 1.

Figure 1. Trajectory data distribution in Fuzhou city

The trajectory data are sparse because of their high collection cost. In addition, different trajectory data may appear in the same driving path due to their asynchronous nature, position shift, and loss of packet transmission. The chaotic and noisy nature of a driving environment could also produce trajectory data that lose valuable information, duplicate existing data, or even cause inconsistent trajectory data. Consequently, trajectory data are not only voluminous but also full of noise and complex features. Because these biased trajectory data are miscellaneous and full of noise, it is difficult to use them as inputs of model data with regularity.

From the recorded trajectory database, four datasets from four vehicles are randomly selected to demonstrate traveling patterns in Figure 2. The results show that these trajectory data are also full of latent patterns and that different drivers have their own unique traveling patterns. It is, therefore, challenging to decompose trajectory data into principal topics and analyze drivers' intentions.



(a) Vehicle 1 (b) Vehicle 2

(c) Vehicle 3 (d) Vehicle 4

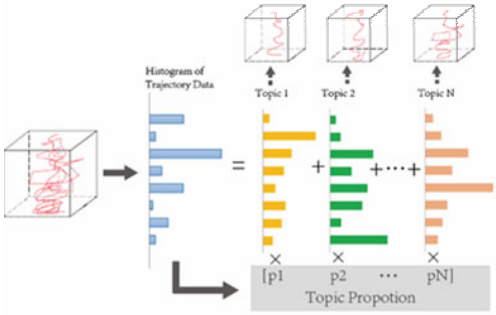
Figure 2. Four examples of trajectory datasets from four vehicles randomly selected in the database

4 Proposed Method

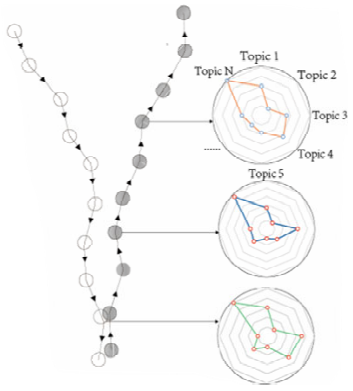
To address the above challenge, a novel TTM based on the unsupervised method was proposed to extract and discriminate drivers' behaviors from massive trajectory data in the semantic level. The TTM decomposes naturalistic trajectory data into a variety of topics and denotes traveling events as topics with certain probabilities.

The TTM is mainly made up of two parts: trajectory representation and trajectory topic modeling. As illuminated in Figure 3(a), naturalistic trajectories are converted into discrete representations with histograms in the first step; then the trajectory topics with different probabilities are extracted in the semantic level.

The topics are used to predict drivers' traveling destinations with the probability distribution of topics. With the trajectory topics, driving behaviors are expressed as probability distributions of the topics. These trajectory topics are then used to quantitatively analyze drivers' general preferences. In addition, the probability distribution is bounded by spatial position, and this location-related distribution of topics could be used in various applications. As shown in Figure 3(b), based on the known spatial location in the TTM, a driver's trip destination can be predicted in real time. The next two sections (A and B) detail the trajectory representation process and the trajectory topic modeling procedure.



(a) Generating procedure of trajectory topic modeling



(b) Application demonstration of trajectory topics

Figure 3. The processing framework and application illustration for the TTM

4.1 Trajectory Representation

Trajectory data are generated from an individual vehicle's movement history. Vehicles equipped with positioning devices periodically reported their location and time stamp, which together generate the vehicle's trajectory data, shown as follows:

$$TR_i = < \rho_{(i,1)}, \cdots, \rho_{(i,j)}, \cdots, \rho_{(i,N_i)}, \cdots > \quad \textbf{(1)}$$

where TR_i is the trajectory of the i^{th} vehicle, n is its total number of reported locations, and p is the location information including longitude, latitude, and time stamp. A vehicle's trajectory is calculated based on a series of traveling events. As expected, a driver's trajectory data always include a large number of traveling events. We set the thresholds (δ_t ,δ_d) as the maximum limit of the grid staying time (the difference in time between when a driver goes into and then leaves a grid) and the hop distance (the spatial distance between the grids in succession), respectively. Then trajectories are separated to two parts as follows:

$$p_i(x_i, y_i, t_i) \in \begin{cases} TR_m & if\ t_i - t_{i-1} < \delta_t \ \&\& \| dist(p_i, p_{i-1}) < \delta_d \| \\ TR_{m+1} & else \end{cases} \quad \textbf{(2)}$$

where $dist(p_i, p_{i-1})$ is the distance between p_i and p_{i-1}.

Trajectory data are represented as a series of spatial grids to reduce the data dimension. First, the whole geographical space was divided into grids. This process is expressed as

$$G = \bigcup_{i=1}^{m*m} g_i \quad \textbf{(3)}$$

where G is the whole geographical space, and g_i is the i^{th} spatial grid.

Then the position of a vehicle is mapped to spatial grids. Consequently, a trajectory is expressed as a vector of grids. That is,

$$TR_i = < g_{(i,1)}, \cdots, g_{(i,2)}, \cdots, g_{(i,n)} > \quad \textbf{(4)}$$

To eliminate duplicated data, the same grids in succession are represented as unique ones. As illuminated in Figure 4, a trajectory is represented as a vector of grids. The trajectory is further reconstructed as a line connecting the middle of successive grids.

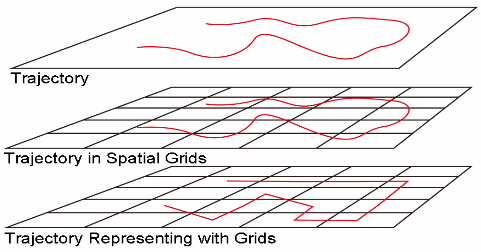


Figure 4. Trajectory representation with spatial grid

Like a bag-of-word model [42] in text mining, we count the staying time of each grid to evaluate the drivers' interesting degrees of grid, which is a measure that evaluates the probability that a driver would drop into a particular grid. Let $\hat{g}_{(i,j)}$ be the grid staying time in grid j of trajectory i and $p_{(i,n)}$ be the n-th data in trajectory i. $p_{(i,n)}.t$ denotes its time stamp. Thus, the grid staying time can be computed as follows:

$$\hat{g}(i,j) = \sum_{n=1}^{N} (p_{(i,n+1)} \cdot t - p_{(i,n)} \cdot t) \quad \textbf{(5)}$$

where $[p_{(j,n+1)}, p_{(j,n)}] \in g_j$ and $N = len(TR_i)$.

4.2 Trajectory Topic Modeling

Let θ_m represent the proportion of each underlying topic for the trajectory TR_m ; $z_{(m,n)}$ indicates the underlying topic proportion of each spatial grid in trajectory TR_m ; $g_{(m,n)}$ denotes the spatial grid vector of trajectory TR_m ; M is the total number of trajectories in the trajectories dataset, and N_m represents the total number of grids in the trajectory; K is the total number of topics. Thus, the trajectory topic model with LDA can be demonstrated as shown in Figure 5.

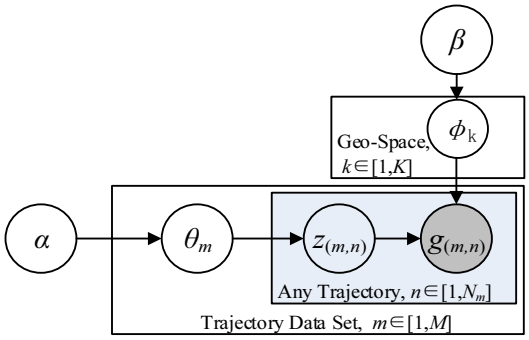


Figure 5. Trajectory topic model with LDA

In the TTM, (α, β) are the super-parameters of the Dirichlet distribution in θ and ϕ . As they are prior parameters set in advance, a small scalar value is generally set. α reflects the density of driving behavior topics in the trajectory data; β represents the density of spatial grids in topics. A greater value for α means a greater number of topics in trajectories, and a greater value of β means more spatial grids in each driving topic.

With the TTM, any driving trajectory could be regarded as a mixed distribution of trajectory topics. Let $D = \{Tr_i\}_{i=1}^{m}$ denote the whole trajectory dataset composed of m trajectories; $Tr_i = \{\hat{g}_{(i,j)}\}_{j=1}^{N_1}$ denotes the vector of grid staying time in the i-th trajectory. Then all explicit and implicit variables in the trajectory topic model are shown as the joint probability

distribution in the following formula:

$$p(g_m, z_m, \theta_m, \Phi \mid \alpha, \beta) = \overbrace{\underbrace{\prod_{n=1}^{N_m} p(g_{m,n} \mid \varphi_{z_{m,n}}) \cdot p(z_{m,n} \mid \theta_m)}_{Spatial\ Grid} \cdot p(\theta_m \mid \alpha) \cdot \underbrace{p(\Phi \mid \beta)}_{Driving\ Topic}}^{Driving\ Trajectory} \tag{6}$$

where $\Phi = \{\phi_k\}_{k=1}^{K}$ is the set of trajectory topics; the probability of $g_{(m,n)}$ can be estimated by the grid staying time.

Then Markov chain Monte Carlo (MCMC) and Gibbs sampling [43] are employed to calculate the other implicit variables such as θ and ϕ. The Gibbs sampling algorithm performs iterative processing on each dimension, and each iteration process chiefly samples one of the dimensions with the others, constrained until the trajectory topic parameters θ, ϕ, and the topic set $\{z_{(m,n)}\}_{i=1}^{K}$ is completely derived.

The trajectory topics are composed of a series of spatial grids with different significant probabilities. With ϕ derived, the probabilities of the grid belonging to any trajectory topic are available. As formula (7) shows, the significant spatial grids were collected to represent trajectory topics with an adjustable threshold that is set as a mean value in this paper:

$$TP_{(i,k)} = \{g_{(i,j)} \mid \phi_k(g_{(i,j)}) > mean(\phi_k)\}_{j=1}^{N} \tag{7}$$

where $TP_{(i,k)}$ is the k-th trajectory topic of driver i; $g_{(i,j)}$ is the j-th spatial grid of driver i, and $\phi_k(g_{(i,j)})$ is its probability belonging to the k-th trajectory topic.

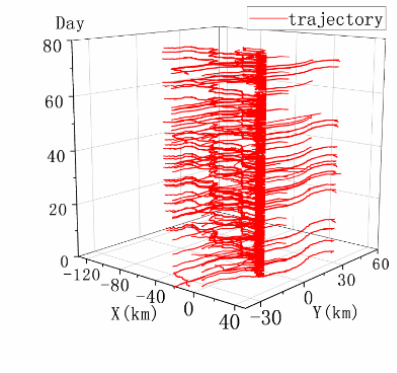
The processing procedure used to discriminate drivers' behavior patterns from trajectory data is dynamic. Prior knowledge of the TTM, presented as parameters (α, β), is updated online by the input of new sample data. With the elapse of time, the drivers' topics patterns would be adjusted by adding new interesting topics and deleting outdated topics that have not been relevant for more than 1 month.

5 Experiments and Results

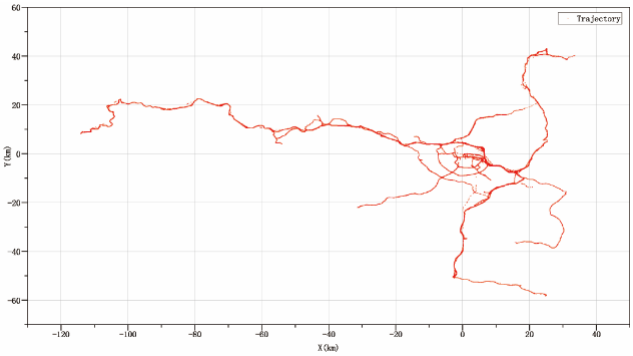
In this section, a number of experiments were conducted to evaluate the feasibility and efficiency of our developed approach using a real trajectory dataset. It should be noted that trajectory data for different types of vehicle have diverse behavior features. For example, the trajectory of buses seems to have a relatively stationary OD intention, but the trajectory of taxis driven by passengers' OD intentions is unpredictable. Therefore, in the following experiments, trajectory data with different types of vehicles were randomly selected to explore drivers' behavior patterns.

5.1 Data Preprocessing

First, the trajectory data are preprocessed. This preprocessing work mainly includes cleaning data, dividing sub-trajectories, and interpolating data. The experiment trajectory dataset in this study is supported by the Fujian Provincial Department of Transportation in China. The dataset comprises data from more than 150,000 commercial vehicles. The experiment data example is shown in Figure 6.



(a) 3-D plot



(b) 2-D plot

Figure 6. Visualization of experimental trajectory data

The first step of preprocessing is to clean abnormal data. Besides physical errors resulting from the transfer of data, other issues concerning complex traffic scenarios also result in dirty trajectory data. For example, the intermittent availability of communication links leads to some leaping steps in driving trajectories data. With the massive amount of data, we detected outliers via the Chebyshev differential equation, which is represented as

$$P(|x - \mu| > k\sigma) \le 1/k^2 \tag{9}$$

The second step is to divide trajectory data into segmentations based on traveling events. The edges between traveling events are clear enough to distinguish from spatial variations and the time intervals between data in succession. The time intervals are particularly key features that express staying points that end the previous traveling event and

start another one. Therefore, in this paper, time intervals are used to distinguish traveling events from trajectory data and to segment trajectories into a series of sub-trajectories for different events.

The third step is to supplement trajectory data with interpolation. Particularly, those trajectory data with much sparsity in nature need to be interpolated. The interpolation in this study aims to make drivers' passing grids continually connected rather than make them smoothly connected. As such, the linear method was adopted to interpolate trajectory data and to adjust the accuracy of the connectivity between grids.

5.2 Trajectory Topic Analysis

Firstly, the spatiotemporal trajectory data is represented as grid vectors. Grid representation was employed to reduce dimension of trajectory data, and each grid size was set to about 100 m. For example, the location (119.476529112, 26.739764512) was represented as grid id 11947626739.

To avoid the influence of road conditions, the repetition grid in the trajectory vector was deleted. When a vehicle keeps stopping in place due to a traffic jam, this similar position information was converted to the same grid id. Note that these duplicate grids would interfere with the analysis of users' behaviors. Thus, it is important to delete duplicate grids to produce successive unique grids.

Then a term frequency matrix with the vector representation of trajectory data was constructed. The grid-represented trajectory data was divided into sub-trajectories by the time interval that is larger than the threshold. Further, we constructed a term frequency matrix, in which the rows correspond to spatial grids, columns correspond to sub-trajectories, and the value of an element is the frequency at which the corresponding grid is visited in the sub-trajectory.

With the matrix of grid frequency, the LDA model was employed to analyze drivers' behavior topics from massive trajectory data. In the model, both α and β were assigned with the small number (0.0002 in this study); the topic number was 8, and the iteration was 20000. The model outputs a topic-grid matrix and a trajectory-topic matrix.

Trajectory topics present a specific path of traveling events. As shown in Figure 7, each topic presents a clear correlation between spatial grids, and when a driver travels through a spatial grid, he/she will usually pass through the other grids in the same topic, so the trajectory topics usually indicate specific traveling events, such as the western activity destination in topic 1, the eastern in topic 4, the southern in topic 7, and the place of the driver's residence in topic 3.

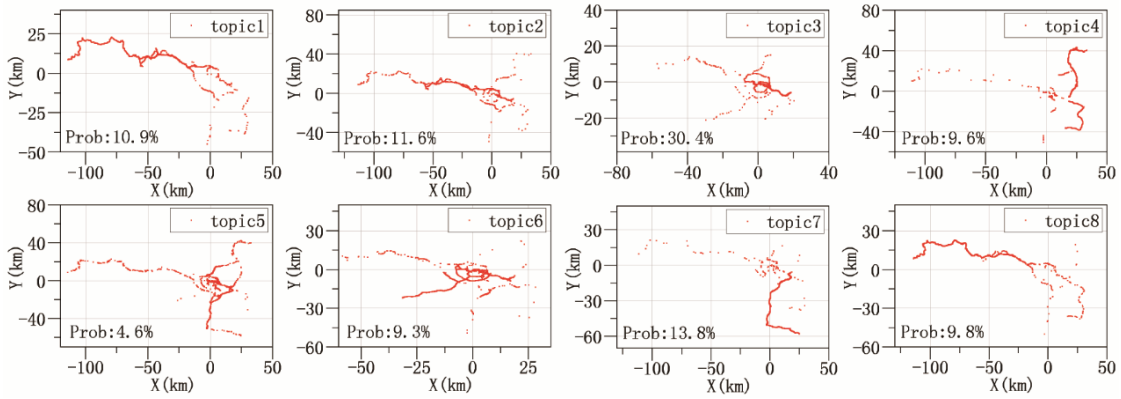


Figure 7. Visualization of trajectory topics

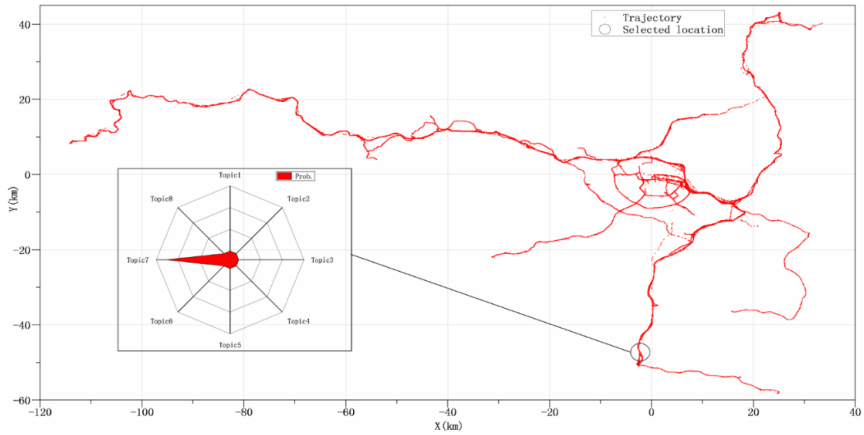


Figure 8. The illustration for topics probability distribution of spatial location

To assess the trajectory topic, the results were also analyzed in two scenarios, including the estimation of driving behavior and the analysis of the drivers' preferences.

First, the trajectory topic could be employed to analyze traveling trajectories with the probability distribution of topics. Figure 9 shows the heat map of probability distribution between traveling trajectories and topics. The gradated color shading represents the correlation between topic and trajectory.

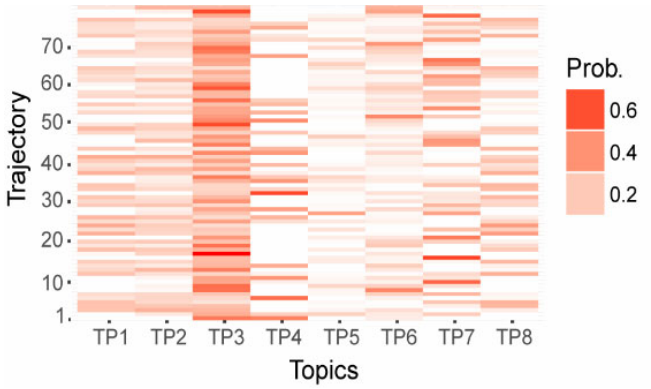


Figure 9. Heat map of correlation between trajectories and topics

From the heat map, it is clear that each trajectory has its own probability distribution of topics and that each topic has a different probability summation. The different probability summations of topics show drivers' traveling preferences. The most likely topic in Figure 9 is topic 3, which demonstrates that the driver prefers to travel in his/her regions, a finding confirmed by his/her residence in this study.

Further, the trajectory topic model can express drivers' intentions, such as driving destinations. The topics-grids matrix outputted from the model represents the topic probability distribution of grids, and the topics probability distribution is determinate if the location is selected. As shown in Figure 8, the driver is determined with most likely to travel in places of topic 7.

5.3 Discussion

The trajectory topic shows the co-occurrence between grids, which is useful in predicting destination and analyzing driving behavior. When the drivers travel through a spatial grid, the probability of traveling to other grids could be estimated by the co-occurrence. In contrast with the existing method of predicting destination and clustering trajectory, the trajectory topics make drivers' behaviors quantitatively analyzable without heavy computation.

On the other hand, the TTM generates topics automatically, which could be viewed as a kind of principal component analysis of trajectory. As shown in Figure 7, the drivers' traveling paths were decomposed into various topics, and any traveling path could be considered as a probability distribution of topics.

Further, drivers' traveling intentions could be analyzed effectively by the trajectory topic model. After the TTM analysis, drivers were assigned to their own topics. The produced topics can be further analyzed to predict the next road segmentation, drivers' traveling preferences, and their current traveling intentions. Clustering topics of all drivers also provides a novel method to analyze the whole transportation network in a city.

While our approach makes it possible to explore drivers' behaviors with fixed topics, it is interesting to dynamically determine the optimal topic quantity from users' trajectory data. This could enable a precise estimation of the preferences of drivers' behaviors. Another important issue is to keep temporal information in the topic model for exploring period patterns. We will consider these issues further in future works.

6 Conclusion

A TTM, an approach to semantically explore drivers' behavior patterns, was proposed to characterize users' behaviors and corresponding intentions using GPS-based driving trajectory data collected from more than 150,000 commercial vehicles in Fujian Province, China.

The TTM allows us to transform a large amount of disorder trajectory data into vectors of regularity and to explore topic probability. This information can help analyze drivers' behaviors from massive trajectory data and demonstrate traveling intentions. Given drivers' sub-trajectory data, the model allows us to analyze patterns that occur frequently. Moreover, the topic probability distribution of grids can help predict the next passing road and even the driver's destination if his/her current location is known.

Acknowledgement

The authors thank Dr. Shengbo Eben Li with the State Key Lab of Automotive Safety and Energy at the Tsinghua University for valuable discussions on this study.

This work was supported in part by Projects of the National Science Foundation of China (41471333, 61304199), project 2017A13025 of Science and Technology Development Center, Ministry of Education, project 2018Y3001 of Fujian Provincial Department of Science and Technology, projects of Fujian Provincial Department of Education (JA14209, JA15325). The Fujian Provincial Department of Transportation is also acknowledged for supporting the experimental dataset.

References

[1] M. Li, X. Li, J. Yin, TORD Problem and Its Solution Based on Big Trajectories Data, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 6, pp. 1666-1677, June, 2016.

[2] G. R. Jagadeesh, T. Srikanthan, K. H. Quek, Heuristic Techniques for Accelerating Hierarchical Routing on Road Networks, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 3, No. 4, pp. 301-309, December, 2002.

[3] S. Wang, Z. Yan, G. Geng, Y. Zhang, Geo-based Content Naming and Forwarding Mechanism for Vehicular Networking over CCN, *International Journal of Internet Technology and Secured Transactions*, Vol. 6, No. 4, pp. 291-302, January, 2016.

[4] Y.-Y. Tseng, J. Knockaert, E. T. Verhoef, A Revealed-preference Study of Behavioural Impacts of Real-time Traffic Information, *Transportation Research Part C: Emerging Technologies*, Vol. 30, pp. 196-209, May, 2013.

[5] W. J. Schakel, B. Van Arem, Improving Traffic Flow Efficiency by In-car Advice on Lane, Speed, and Headway, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 4, pp. 1597-1606, August, 2014.

[6] Y. Deguchi, K. Kuroda, M. Shouji, T. Kawabe, HEv Charge/Discharge Control System Based on Navigation Information, *International Congress, Transportation Electronics, Vehicle Electronics to Digital Mobility: The Next Generation of Convergence*, Detroit, MI, 2004, pp. 217-224.

[7] H. Tan, Y. Wu, B. Shen, P. J. Jin, B. Ran, Short-term Traffic Prediction Based on Dynamic Tensor Completion, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 8, pp. 2123-2133, August, 2016.

[8] Y. Lv, Y. Duan, W. Kang, Z. Li, F.-Y. Wang, Traffic Flow Prediction with Big Data: A Deep Learning Approach, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 2, pp. 865-873, April, 2015.

[9] E. I. Vlahogianni, M. G. Karlaftis, J. C. Golias, Short-term Traffic Forecasting: Where We Are and Where We're Going, *Transportation Research Part C: Emerging Technologies*, Vol. 43, Part 1, pp. 3-19, June, 2014.

[10] A. Abadi, T. Rajabioun, P. A. Ioannou, Traffic Flow Prediction for Road Transportation Networks with Limited Traffic Data, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 2, pp. 653-662, April, 2015.

[11] E. I. Vlahogianni, B. B. Park, J. Van Lint, Big Data in Transportation and Traffic Engineering, *Transportation Research Part C: Emerging Technologies*, Vol. 58, Part B, pp. 161, September, 2015.

[12] L. Wang, Z. Yu, B. Guo, T. Ku, F. Yi, Moving Destination Prediction Using Sparse Dataset: A Mobility Gradient Descent Approach, *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 11, No. 3, Article No. 37, April, 2017.

[13] T. Gindele, S. Brechtel, R. Dillmann, A Probabilistic Model for Estimating Driver Behaviors and Vehicle Trajectories in Traffic Environments, *13th International IEEE Conference on Intelligent Transportation Systems*, Funchal, Madeira Island, Portugal, 2010, pp. 1625-1631.

[14] T. Hunter, P. Abbeel, A. Bayen, The Path Inference Filter: Model-based Low-latency Map Matching of Probe Vehicle Data, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 2, pp. 507-529, April, 2014.

[15] S. Qiao, N. Han, W. Zhu, L. A. Gutierrez, TraPlan: An Effective Three-in-One Trajectory-Prediction Model in Transportation Networks, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 3, pp. 1188-1198, June, 2015.

[16] V. S. Tiwari, S. Chaturvedi, A. Arya, Route Prediction Using Trip Observations and Map Matching, *3rd IEEE International Advance Computing Conference (IACC)*, Ghaziabad, India, 2013, pp. 583-587.

[17] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, V. S. Tseng, Semantic Trajectory Mining for Location Prediction, *19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Chicago, Illinois, 2011, pp. 34-43.

[18] L. Zhang, X. Sun, H. Zhuge, Location-Driven Geographical Topic Discovery, *Ninth International Conference on Semantics, Knowledge and Grids*, Beijing, China, 2010, pp. 210-213.

[19] E. M. Carboni, V. Bogorny, Inferring Drivers Behavior through Trajectory Analysis, In: P. Angelov, K. T. Atanassov, L. Doukovska, M. Hadjiski, V. Jotsov, J. Kacprzyk, N. Kasabov, S. Sotirov, E. Szmidt, S. Zadrozny (Eds.), *Intelligent Systems' 2014*, Advances in Intelligent Systems and Computing, Vol. 322, pp. 837-848, Springer, 2015.

[20] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, January, 2003.

[21] T. Gindele, S. Brechtel, R. Dillmann, Learning Driver Behavior Models from Traffic Observations for Decision Making and Planning, *IEEE Intelligent Transportation Systems Magazine*, Vol. 7, No. 1, pp. 69-79, January, 2015.

[22] T. Hunter, P. Abbeel, A. M. Bayen, The Path Inference Filter: Model-based Low-latency Map Matching of Probe Vehicle Data, In: E. Frazzoli, T. Lozano-Perez, N. Roy, D. Rus (Eds.), *Algorithmic Foundations of Robotics X*, *Proceedings of the Tenth Workshop on the Algorithmic Foundations of Robotics*, pp. 591-607, Springer-Verlag Berlin Heidelberg, 2013.

[23] J. C. Niebles, H. Wang, F.-F. Li, Unsupervised Learning of Human Action Categories Using Spatial-temporal Words, *International Journal of Computer Vision*, Vol. 79, No. 3, pp. 299-318, September, 2008.

[24] T. A. Arentze, Adaptive Personalized Travel Information Systems: A Bayesian Method to Learn Users' Personal Preferences in Multimodal Transport Networks, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 14, No. 4, pp. 1957-1966, December, 2013.

[25] J.-M. Kim, H. Baek, Y.-T. Park, *Probabilistic Graphical Model based Personal Route Prediction in Mobile Environment, Applied Mathematics and Information Sciences*, Vol. 6, No. 2S, pp. 651S-659S, April, 2012.

[26] K. Chen, J.-K. Kämäräinen, Pedestrian Density Analysis in Public Scenes With Spatiotemporal Tensor Features, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 7, pp. 1968-1977, July, 2016.

[27] T. Rattenbury, N. Good, M. Naaman, Towards Automatic Extraction of Event and Place Semantics from Flickr Tags, *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Amsterdam, Netherlands, 2007, pp. 103-110.

[28] I. Dagli, M. Brost, G. Breuel, Action Recognition and Prediction for Driver Assistance Systems Using Dynamic Belief Networks, *International Conference on Object-Oriented and Internet-Based Technologies, Concepts, and Applications for a Networked World*, Erfurt, Germany, 2002, pp. 179-194.

[29] I. Dagli, D. Reichardt, Motivation-based Approach to Behavior Prediction, *IEEE Intelligent Vehicle Symposium*, Versailles, France, 2002, pp. 227-233.

[30] F. Castaldo, F. A. N. Palmieri, V. Bastani, L. Marcenaro, C. Regazzoni, Abnormal Vessel Behavior Detection in Port Areas based on Dynamic Bayesian Networks, *17th International Conference on Information Fusion*, Salamanca, Spain, 2014, pp. 1-7.

[31] P. Liu, A. Kurt, U. Ozguner, Trajectory Prediction of a Lane Changing Vehicle based on Driver Behavior Estimation and Classification, *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, Qingdao, Shandong, China, 2014, pp. 942-947.

[32] R. M. Paven, M. Pachia, D. Pescaru, A Driver Behavior Learning Framework for Enhancing Traffic Simulation, *Carpathian Journal of Electronic and Computer Engineering*, Vol. 7, No. 1, pp. 7-12, June, 2014.

[33] J. Park, Y. L. Murphey, R. McGee, J. G. Kristinsson, M. L. Kuang, A. M. Phillips, Intelligent Trip Modeling for the Prediction of an Origin–destination Traveling Speed Profile, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 3, pp. 1039-1053, June, 2014.

[34] M. Brannstrom, E. Coelingh, J. Sjoberg, Model-based Threat Assessment for Avoiding Arbitrary Vehicle Collisions, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 11, No. 3, pp. 658-669, September, 2010.

[35] P. Lytrivis, G. Thomaidis, A. Amditis, Cooperative Path Prediction in Vehicular Environments, *11th International IEEE Conference on Intelligent Transportation Systems*, Beijing, China, 2008, pp. 803-808.

[36] Y. Liu, M. Ester, B. Hu, D. W. Cheung, Spatio-Temporal Topic Models for Check-in Data, *IEEE International Conference on Data Mining*, Atlantic, NJ, 2015, pp. 889-894.

[37] S. Chiappino, P. Morerio, L. Marcenaro, C. S. Regazzoni, A Bio-inspired Knowledge Representation Method for Anomaly Detection in Cognitive Video Surveillance Systems, *16th International Conference on Information Fusion*, Istanbul, Turkey, 2013, pp. 242-249.

[38] J. Cao, K. Zeng, H. Wang, J. Cheng, F. Qiao, D. Wen, Y. Gao, Web-based Traffic Sentiment Analysis: Methods and Applications, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 2, pp. 844-853, April, 2014.

[39] A. Bender, G. Agamennoni, J. R. Ward, S. Worrall, E. M. Nebot, An Unsupervised Approach for Inferring Driver Behavior from Naturalistic Driving Data, *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 6, pp. 3325-3336, December, 2015.

[40] Z. Yin, L. Cao, J. Han, C. Zhai, T. Huang, Lpta: A Probabilistic Model for Latent Periodic Topic Analysis, *IEEE 11th International Conference on Data Mining*, Vancouver, BC, Canada, 2011, pp. 904-913.

[41] S. Lefèvre, D. Vasquez, C. Laugier, A Survey on Motion Prediction and Risk Assessment for Intelligent Vehicles, *Robomech Journal*, Vol. 1, No. 1, p. 1, December, 2014.

[42] G. Salton, A. Wong, C.-S. Yang, A Vector Space Model for Automatic Indexing, *Communications of the ACM*, Vol. 18, No. 11, pp. 613-620, November, 1975.

[43] G. Heinrich, *Parameter Estimation for Text Analysis*, Technical report, May, 2005.

Biographies

Lyuchao Liao received his B.E. and M.S. degree in Information Science from Fuzhou University, and received his Ph.D. degree in Traffic Information Engineering and its Control from Central-South University in 2015. Now he works as a postdoctoral researcher in Tsinghua University, and his research interest is primarily in the field of big data analysis on transportation domain. In particular, he is interested in trajectory data mining and its applications, such as driving behavior analysis, traffic state prediction and traffic road-network optimization.

Jianping Wu graduated from Zhejiang University in 1982 and received his Ph.D. degree in Transportation Engineering from University of Southampton, UK, in 1994. He is a professor in Department of Civil Engineering at Tsinghua University and the recipient of the prestigious "Chong Kong Scholar Professorship" awarded by the Ministry of Education of China. He is also a Visiting Professor at the University of Southampton, UK.. His research interest is primarily in the field of sustainable transport systems and low carbon transport, microscopic transport simulation, and intelligent transport systems.

Fumin Zou received the Ph.D. degree in Traffic Information Engineering & Control from Central-South University, Changsha, China, in 2009. He visited Texas Tech University in Dec. 2013 – Dec. 2014. He is a Professor with the School of Information Science and Engineering, Fujian University of Technology (FJUT) and the Innovation Center of Beidou Navigation and Smart Traffic of Fujian Province (ICBNST), China. His current research interesting includes machine learning, big data and the traffic information system.

Jeng-Shyang Pan received the B.S. degree in Electronic Engineering from the National Taiwan University of Science and Technology in 1986, the M.S. degree in Communication Engineering from the National Chiao Tung University, Taiwan in 1988, and the Ph.D. degree in Electrical Engineering from the University of Edinburgh, U.K. in 1996. Currently, he is a Dean for College of Information Science and Engineering, Fujian University of Technology and a director of Innovative Information Industry Research Center, Harbin Institute of Technology Shenzhen Graduate School, China. He jointed the editorial board of LNCS Transactions on Data Hiding and Multimedia Security, Journal of Computers, Journal of Information Hiding and Multimedia Signal Processing etc. His current research interests include soft computing, information security and signal processing.

Tingting Li received B.E. degree in transportation engineering from Beijing Jiaotong University, Beijing, China in 2015. She is currently working toward the Ph.D. degree with the department of civil engineering. Tsinghua University, China. Her research interests include traffic big data, machine learning and the applications in autonomous vehicles.