TourSense: A Framework for Tourist Identification and Analytics Using Transport Data

Yu Lu, Huayu Wu, Xin Liu, and Penghe Chen

Abstract—We advocate for and present *TourSense*, a framework for tourist identification and preference analytics using city-scale transport data (bus, subway, etc.). Our work is motivated by the observed limitations of utilizing traditional data sources (e.g., social media data and survey data) that commonly suffer from the limited coverage of tourist population and unpredictable information delay. *TourSense* demonstrates how the transport data can overcome these limitations and provide better insights for different stakeholders, typically including tour agencies, transport operators, and tourists themselves. Specifically, we first propose a graph-based iterative propagation learning algorithm to recognize tourists from public commuters. Taking advantage of the trace data from the identified tourists, we then design a tourist preference analytics model to learn and predict their next tour, where an interactive user interface is implemented to ease the information access and gain the insights from the analytics results. Experiments with real-world datasets (from over 5.1 million commuters and their 462 million trips) show the promise and effectiveness of the proposed framework: the Macro and Micro F1 scores of the tourist identification system achieve 0.8549 and 0.7154, respectively, whereas the tourist preference analytics system improves the baselines by at least 23.53 and 11.44 percent in terms of precision and recall.

Index Terms—Data mining and knowledge discovery, transportation systems, tourist recommendation

1 Introduction

As one of the world's largest industries, tourism serves as the economic pillar of many countries and cities. The total contribution of the tourism industry to GDP was 7,600 billion U.S. dollars (3.1 percent of global GDP) and supported 292 million jobs (9.6 percent of total employment) in 2016.[1] Taking Singapore as an example, its tourism industry brought in more than 16.4 million of foreign tourists (more than thrice the country's population) and created more than 160 thousand jobs for local residents in 2017. Tracking and understanding tourists would directly benefit local government and tour agencies to design and improve their services, such as launching new tour routes and providing customized tour packages based on tourist's characteristics and preferences.

To capture and understand tourists and their preferences, the recent tourism analytics research mainly adopts social media data (e.g., geotagged images in Flickr) [1], [2], [3], where the basic assumption behind this attempt is that most tourists would like to share their travel moments on their online social networks. However, using social media data may suffer from the *limited coverage* and *information delay*: (a) only a small portion of tourists are actively sharing their photos or travel experiences on social media, as many travellers may not be the fans of social networks or even not use the Internet. Furthermore, most shared contents are popular landmarks, not covering all the places a tourist visited, and thus the insight gained from social media data may be incomplete or biased; (b) considering the high data roaming fees, many social network sharings are not real-time posted. Tourists may share their photos and feelings after a whole day's travel, or even after coming back to their hometowns. Meanwhile, how to effectively and timely crawl all the tourists' social media information from the service providers is also challenging. Besides the social media data, sensor network data (e.g., bluetooth data) [4] and cellular data [5] are also adopted by the researchers for tourist study, but they suffer from the similar limitations and constraints.

This work attempts to tackle the above issues, by demonstrating how the transport data can be used to identify and analyze tourists. Despite of a diversity of local tour services available, public transport (e.g., metro and bus) is still the most cost-efficient and convenient travelling approach for most tourists, especially in the densely-populated cities like Singapore and Tokyo. Accordingly, the public transport data offer a sufficient coverage of the tourist population. Meanwhile, the widely adopted electronic fare payment systems can timely record and trace tourists and their travelling routes, when they tap in/ out at the gantry of a station or boarding/alighting on a bus. In particular, we propose a novel but practical framework for tourist analytics,

1. https://www.wttc.org/-/media/files/reports/economic-impact-research/regions-2017/world2017.pdf

- *Y. Lu and P. Chen are with the Advanced Innovation Center for Future Education, School of Educational Technology, Faculty of Education, Beijing Normal University, Beijing 100875, China. E-mail: {luyu, chenpenghe}@bnu.edu.cn.*
- *H. Wu is with Nanyang Technological University, 639798, Singapore. E-mail: wu_huayu@ntu.edu.sg.*
- *X. Liu is with the School of Automation, Hangzhou Dianzi University, Hangzhou 310002, China. E-mail: liu-x@i2r.a-star.edu.sg.*

Manuscript received 1 June 2018; revised 2 Jan. 2019; accepted 8 Jan. 2019. Date of publication 21 Jan. 2019; date of current version 4 Nov. 2019. (Corresponding author: Penghe Chen.) Recommended for acceptance by J. Xu. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TKDE.2019.2894131

called *TourSense*, that (a) first applies machine learning techniques on transport data to identify tourists from public commuters, and (b) uses the identified tourist travelling information to conduct their preference analytics and thereby timely makes the personalized recommendation and prediction. To provide the practical embodiments of the proposed framework, we take Singapore as an exemplary case and present the empirical experiment results using the public transport data from the city.

Our work in this paper thus makes the following *key contributions*:

- *Novel Framework for Tourist Analytics*: We propose a novel framework that conducts analytics on tourists using transport data. By leveraging on the citywide bus and subway data, we show how the public transport data can provide *hard-to-obtain*, *tourist-specific* insights and quantitative results.
- *Identification on Tourists from Public Commuters*: Using the transport data, we propose a two-phase algorithm to identify tourists from public commuters. The key innovations include (i) properly ranking transport stations according to how they are likely to be a destination for tourists; and (ii) designing a graph-based novel iterative learning algorithm to accomplish the tourist identification.
- *Tourist Preference Analytics*: Using the identified tourists and their travel records, we design the personalized preference analytics and location recommendation methods for tourists. The key innovation include (i) a tourist-location transition frequency matrix and a location-location transition frequency matrix are designed to represent the tourist information, and (ii) a novel recommendation model is designed to learn tourists' preferences for individual locations and tours. To the best of our knowledge, this is the first work that analyzes tourists' public transport trajectories for location preference study.
- *Real-World Experiment and Comprehensive Evaluation*: Using the real world data from 5.1 million public commuters and their 462 million trips, we have conducted the comprehensive evaluations, which show that the proposed framework can identify the tourist with a F1 score over 0.85 and meanwhile outperform all the four baselines on the personalized location recommendation by at least 23.53 and 11.44 percent in terms of precision and recall.

While our tourist identification and preference analytics are both novel, we believe that the key impact of this work is to highlight a broader possibility of understanding tourists, and accordingly create innovative and personalized services for tourists, based on the novel data sources and information systems.

The rest of the paper is organized as follows. Section 2 depicts the overall framework architecture. Section 3 describes the tourist identification system design. Section 4 presents the tourist preference analytics system design. We show the experimental results and demonstrate the user interface in Section 5. The discussion and related work are then given in Sections 6 and 7. Finally, we conclude in Section 8.

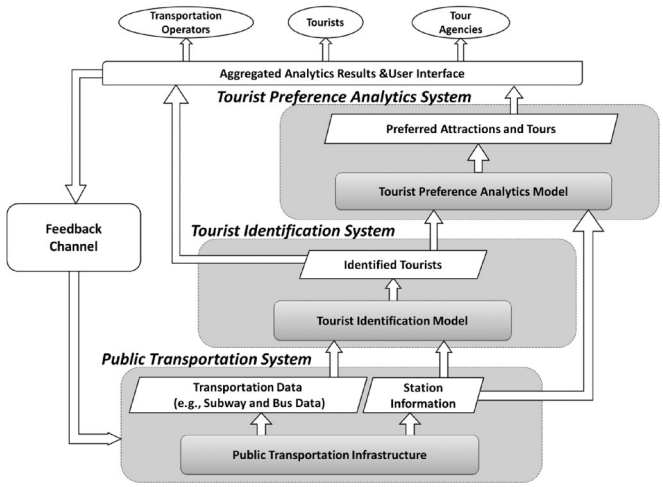


Fig. 1. Block diagram of the *TourSense* framework.

2 System Overview

The block diagram of the *TourSense* framework is illustrated in Fig. 1, which mainly consists of three modules, namely *public transportation system*, *tourist identification system* and *tourist preference analytics system*. Briefly speaking, *public transportation system* provides the transportation data and infrastructure information (e.g., subway/bus data and station information). By leveraging on such data and information, *tourist identification system* recognizes tourists from public commuters. Using the identified tourists and their travelling traces, *tourist preference analytics system* further investigates their favorite attractions and tours. All the above tourist information and analytics results will be aggregated, via specifically designed user interface and feedback channel, and eventually provide to different stakeholders, typically including transportation operators, government agencies and tourists themselves. We will elaborate the three systems in this section respectively.

2.1 Public Transportation System

The public transportation infrastructure covers different urban transportation services (e.g., subway and bus services) and facilities (e.g., subway and bus stations). Each service utilizes their own informatics system to acquire relevant commuter travelling data. For example, today's subway service usually employs RFID-based ticketing system to automatically collect passengers' ingress and egress (tap-in and tap-out) data at each subway station. The bus service deploys the similar ticketing system on each operating bus to record their boarding and alighting information. The commuter travelling data include both real time transaction data, which can be timely collected by the backend servers of transportation operators, and the historical transaction data, which are usually streamed and stored into a Hadoop distributed file system for daily maintenance and batch processing. In addition, the public transportation system will also provide the station information, including the geographic location and nearby POIs of each metro station or bus stop.

2.2 Tourist Identification System

This system periodically recognizes tourists from commuters using the data and information collected from the public

transportation system. More specifically, it targets to identify the transport records that are generated by the riding of tourists from the public transport data. In general, the travelling population can be assumed as two groups, i.e., tourists and non-tourists (non-tourists normally mean local people). Tourists refer to the group of people who visit the city for sightseeing purpose during a short term (e.g., a couple of days). They commonly visit places of interest, including historic sites, museums, restaurants, shopping streets, and stay in hotels or hostels. People who come to the city for other purposes such as business or medical services may not fall into the class of tourists in this system. Some local domain knowledge and a small set of labeled commuters information may be needed during the identification process. The key outputs of the system is the identified tourist sets and their riding records, which serve as the main inputs of the upper tourist preference analytics system.

2.3 Tourist Preference Analytics System

Taking advantages of the identified tourist information, especially their travelling traces, this system mainly conducts the preference analytics on the tourists, such as predicting individual tourist's next visiting locations and accordingly making next POI (place of interest) recommendations to those who are not sure about where to go. Such preference analytics results can be utilized in many services. For example, the inferred tourist preferences on his or her unvisited locations can be used to generate the personalized advertisement (e.g., attraction tickets and nearby dining promotions), which can be pushed to the tourists through different feedback channels, such as the screens on the subway station gantry or the top-up machines at the ticketing office. Moreover, the analytics results can be used by the designed user interface to answer "next-visiting-place" queries from tourists.

In short, the above-described three systems work cooperatively to acquire, process and analyze the public transportation data for tourists. The final analytics results would possibly benefit different stakeholders, including tourists, transportation operators and tour agencies. We will elaborate our design on tourist identification system and preference analytics systems in the subsequent two sections.

3 Tourist Identification System Design

We design a two-phase algorithm to tackle the tourist identification problem. The first phase conducts the so-called *station ranking*. Its main task is to assign an initial score to each transportation station that indicating whether it is more likely a destination for tourists or a destination for non-tourists. The second phase conducts the so-called *iterative propagation learning*, where an iterative learning algorithm is designed using the station ranking results to accomplish the tourist identification task. We will present the two phases in the following parts respectively.

3.1 Phase I: Station Ranking

Intuitively, knowing someone who has visited a station with a high (or low) initial score may increase (or reduce) our belief that the person is a tourist. We thus compute a score for each given station to describe whether the station is more likely to be a destination for tourists. However, it is not a proper way to simply use the attractiveness of a place to tourists as the initial scores (such as the scores on the travel sites like TripAdvisor). It is mainly because one place that is popular to tourists may also be popular to locals. For example, most tourists may visit famous shopping streets in a city, while local people may favorite them as well. We thus need to consider the popularity of a place to both tourists and locals when computing the initial score for each station.

One way to compute the score for a station is using the probability of being a tourist, given that a commuter has visited that station. For simplicity, we simply denote this probability as $\Pr(t|m_i)$, where t denotes that a commuter is a tourist and m_i denotes that the commuter has visited the ith station. Computing the exact value of $\Pr(t|m_i)$ is not a straightforward task, and we thus transform it using Bayes' theorem:

$$\Pr(t|m_i) = \Pr(t) \cdot \frac{\Pr(m_i|t)}{\Pr(m_i)}, \tag{1}$$

where $\Pr(m_i|t)$ is the probability of a commuter to visit the station given that he/she is a tourist, $\Pr(m_i)$ is the prior probability of a commuter to visit the station, and $\Pr(t)$ is the prior probability of a commuter to be a tourist. In the following, we elaborate the estimation on the above three terms respectively.

3.1.1 *Estimation on* $\Pr(m|t)$

Computing $\Pr(m|t)$ needs the data from tourists (not necessarily the entire data from all tourists) to summarize how often they visit each station. It is commonly known that many tourists choose to buy one-time ticket when taking pubic transport, which can be used to conduct the estimation here. Let n_i^t be the number of tourists using one-time tickets at the station ith station, we can estimate the probability using the maximum likelihood principle:

$$\hat{\Pr}(m_i|t) = \frac{n_i^t}{\sum_i n_i^t}, \tag{2}$$

where $\hat{\Pr}(m_i|t)$ is the estimate of the desired probability, $\sum_i n_i^t$ is the total number of the tourists at all the stations. However, one concern is that local commuters may also buy such one-time tickets, especially when they forget to bring their regular ticket (e.g., electronic transport card) or have no sufficient balance inside it. Such cases may be rare, but the total rides from local commuters are much larger than the rides from tourists. Hence, it is necessary to exclude such cases before using one-time ticket transactions. In the experiment section, we will show how to estimate n_i^t using the real-world public transport data.

3.1.2 *Estimation on* $\Pr(m_i)$

$\Pr(m_i)$ describes the overall probability that one may visit the ith station regardless one is a local commuter or a tourist. Using the maximum likelihood principle, we can estimate it as follows:

$$\hat{\Pr}(m_i) = \frac{n_i^s + n_i^r}{\sum_i (n_i^s + n_i^r)}, \tag{3}$$

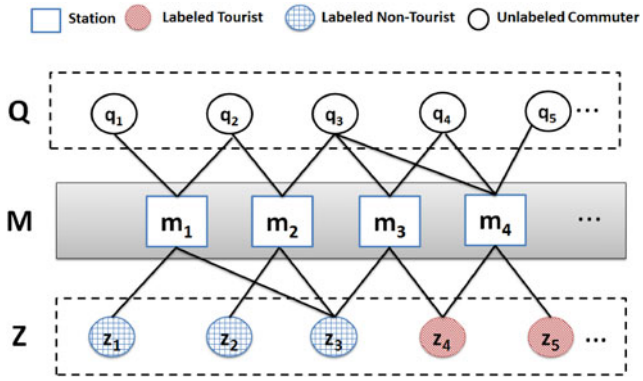


Fig. 2. Station-commuter relationship (SCR) graph.

where $\sum_i (n_i^s + n_i^r)$ is the total number of records from both regular tickets and one-time tickets at all the stations.

3.1.3 Estimation on $\Pr(t)$

$\Pr(t)$ mainly describes how likely a commuter is a tourist, which can be regarded as a constant coefficient for all the stations in Eq. (1). The value of $\Pr(t)$ does not affect the ranking of the station in terms of the assigned scores, and accordingly a coarse estimate on $\Pr(t)$ could be sufficient in practice. Specifically, given n_i^t as the number of tourist visiting the ith station using one-time ticket, we can simply assume the same number of tourist using regular tickets at the same station. Accordingly, an estimate on $\Pr(t)$ can be made as follows:

$$\hat{\Pr}(t) = \frac{\sum_i 2n_i^t}{\sum_i (n_i^s + n_i^r)}, \tag{4}$$

where $\sum_i 2n_i^t$ is the total number of tourists using either regular or one-time tickets at the ith station. In short, using Eqs. (2), (3) and (4), we can obtain how likely a commuter is a tourist given that he or she has visited a station, namely the desired initial score $\Pr(t|m_i)$ for the ith station.

Note that the current estimation on $\Pr(t)$ is based on the simple assumption of the same number of tourists using one-time ticket as regular ticket, which can be adjusted if any further information and prior knowledge introduced from the tourist side.

3.2 Phase II: Iterative Propagation Learning

3.2.1 Station-Commuter Relationship Graph

To accomplish the tourist identification task, we propose a graph structure, called station-commuter relationship (SCR) graph, to encapsulate all the prior knowledge, including the initial scores on all stations, labeled and unlabeled commuters in the data, as well as the interactional relationships between stations and commuters. As illustrated in Fig. 2, three types of nodes are defined in a SCR graph: the top layer Q consisting of nodes that represent the unclassified commuters, the middle layer M consisting of nodes that represent public transport stations, and the bottom layer Z consisting of nodes that represent the labeled commuters (labeled as tourist or non-tourist). The weighted edge between a commuter and a station indicates that the commuter has visited that station, where the weight is the number of visiting times. The stations mainly act as bridges between commuter sets Q and Z, where the labeled commuter set Z is usually much smaller than the unlabeled set Q.

3.2.2 SCR Graph Initialization

For each node in the SCR graph, it carries an initial probability distribution: for each station node $m_i \in M$, its initial probability distribution can be defined as $[\Pr(m_i), 1-\Pr(m_i)]$, where $\Pr(m_i)$ is calculated by Eq. (1). For each labeled commuter node $z_i \in Z$, its initial probability distribution would be either $[1, 0]$ or $[0, 1]$, as it has been labeled as tourist or non-tourist class. For each node $q_i \in Q$, its initial probability distribution can be defined as $[\Pr(q_i), 1-\Pr(q_i)]$, where $\Pr(q_i)$ can be computed by Eq. (4).

3.2.3 Iterative Propagation Learning

After determining and initializing the desired information from commuters and stations in the defined SCR graph, we design a specific algorithm to cast the tourist identification problem into a node-labelling problem. To update the probability distributions of each node $q_i \in Q$ using the existing information contained in the SCR graph, we propose an iterative propagation learning algorithm. The basic idea behind the proposed algorithm is that the commuters who belong to the same class (tourist or non-tourist) tend to visit similar stations and vice versa. Accordingly, the algorithm mainly consists of two steps:

- *STEP 1*: Update the probability distributions of commuters based on the stations they have visited. In each iteration, the probability distributions for all commuter nodes are updated accordingly. Specifically, for the nodes in Q and Z, their probability values are updated based on the values of their adjacent nodes and their own values in the last step of the iteration. In each iteration step, the following updating rules can be used:

$$\phi_{q_i}^k \leftarrow \alpha \cdot \phi_{q_i}^{k-1} + (1-\alpha) \cdot \frac{\sum_{m \in N(q_i)} w_{q_i m} \cdot \phi_m^k}{\sum_{m \in N(q_i)} w_{q_i m}}, \tag{5}$$

$$\phi_{z_i}^k \leftarrow \beta \cdot \phi_{z_i}^{k-1} + (1-\beta) \cdot \frac{\sum_{m \in N(z_i)} w_{z_i m} \cdot \phi_m^k}{\sum_{m \in N(z_i)} w_{z_i m}}, \tag{6}$$

  where $\phi_{q_i}^k$ and $\phi_{z_i}^k$ are the probability distributions of commuter q_i and z_i at the kth step of iteration. $N(q_i)$ and $N(z_i)$ return all stations commuter q_i and z_i have visited respectively, and $w_{q_i m}$ and $w_{z_i m}$ are the weighted edge values, namely the number of times commuters q_i and z_i have visited the station m, respectively. The configurable parameter α and β are used to control the rate of update, where larger values imply greater trust in the probability distributions from the last step of iteration.
- *STEP 2*: Update the probability distributions of stations based on commuters who have travelled to the stations. In each iteration, the probability distributions for all station nodes are updated accordingly. Specifically, for the nodes in M, their probability values are updated based on the commuter nodes and

their own values in the last step of the iteration. In each iteration, the following updating rule can be used:

$$\phi_{m_i}^{k} \leftarrow \gamma \cdot \phi_{m_i}^{k-1} + (1-\gamma) \cdot \frac{\sum_{u \in N(m_i)} w_{um_i} \cdot \phi_{m_i}^{k-1}}{\sum_{u \in N(m_i)} w_{um_i}}, \tag{7}$$

where $\phi_{m_i}^{k}$ is the probability distributions of station m_i at the kth step of iteration. $N(m_i)$ returns all commuters who have visited station m_i, and w_{um_i} is the weighted edge values, namely the number of times commuter u has visited the station m_i. Similar to *STEP 1*, the configurable parameter γ is used to control the rate of update, where a larger value implies greater trust in the probability distribution from the last step of iteration.

Briefly speaking, the algorithm iteratively propagates the tourist and non-tourist information to the unknown commuters using the stations as a bridge. The iterative learning process will end when the number of iterations reaches a predefined threshold or the results converge. After that, each commuter $q_i \in Q$ can be assigned a class label $\hat{C}$ to determine whether she is a tourist or not, such that:

$$\hat{C} = \underset{c}{argmax} \frac{Pr(q_i|c)}{\sum_{q_i \in Q} Pr(q_i|c)}, \tag{8}$$

where c is the class label (either tourist or non-tourist), and $Pr(q_i|c)$ is the corresponding probability. Note that we do not directly use the individual probability distribution results to assign the commuter class, but first perform the normalization using all information from all the unknown commuters in set Q. After that, the class with a larger normalized value will be used to label the commuter.

For each step of the iteration, each node and its neighbors are visited. Hence, the total time complexity of the algorithm is $\mathcal{O}(2k|E|)$, where $|E|$ is the number of edges in the graph and k is the total iteration number. When necessary, the algorithm can be further speeded up by parallel execution of probability distribution update. For example, the update of the probability distributions for each node in Z and each node in Q can be executed concurrently, as their nodes are not adjacent. Furthermore, the update of the probability distributions for the nodes in M can be executed concurrently.

In short, we propose the station ranking and iterative propagation learning algorithms to accomplish the tourist identification tasks, where the SCR graph is used to propagate the knowledge learned from the station information and the labeled commuters.

4 Tourist Preference Analytics System Design

After the tourists are identified from the public commuters, their travel information, especially their travel locations can be directly obtained from their transport riding records. Based on such information, the system conducts the analytics for tourist location preferences. Specifically, our current design is to predict and recommend (1) the next public transport alighting location, i.e., the corresponding attraction that a tourist will visit,[2] and (2) the associated next public transport boarding location, i.e., the end point of the tour. The first one provides the basic personalized recommendation service, while the second one enhances the personalization and the comprehensiveness of the services.

4.1 Model Description

We denote a set of the identified tourists by $U = \{u_1, u_2, ..., u_{|U|}\}$ and a set of locations (i.e., subway stations and bus stops) by $L = \{l_1, l_2, ..., l_{|L|}\}$. Note that $l_i \in L$ contains the descriptive information such as station name, longitude, latitude, etc. For each tourist $u \in U$, her historical transport records, i.e., tours (in chronological order) is denoted by $C_u = \{c_1, c_2, ..., c_n\}$, where each tour $c_i =< l_x, l_y >$ consists of a public transport alighting location l_x and the next boarding location l_y.

We first build the tourist-location visit frequency matrix M_{ul} to record the count of visits between tourists and locations. Intuitively, the higher the visit frequency is, the more the location is preferred by the corresponding tourist. In order to estimate a tourist's ranking preference for locations, we first denote tourist u's preference for location l by a binary variable $\varphi_{u,l}$:

$$\varphi_{u,l} = \begin{cases} 1 & \text{if } r_{u,l} > 0 \\ 0 & \text{if } r_{u,l} = 0 \end{cases},$$

where $r_{u,l}$ indicates the count that u visited l. That is, if a tourist u has visited a location l at least once, we can infer that u likes l, otherwise, u has no explicit preference for l. Intuitively, as the visit count grows, we are confident that the tourist u likes the location l. By considering the visit counts, we define a confidence level $\theta_{u,l}$ for the corresponding preference $\varphi_{u,l}$:

$$\theta_{u,l} = 1 + \alpha^{u} * r_{u,l}, \tag{9}$$

where α^{u} is a variable controlling the influence of the increase of the visit counts.

Accordingly, the target tourist u's preference for the target location l can be estimated by the inner product of the latent factor vectors for each tourist and each location:

$$\hat{\varphi}_{u,l} = \mathbf{u}_u^{\top} \mathbf{v}_l, \tag{10}$$

where $\mathbf{u}_u$ is the latent factor vector of tourist u, and $\mathbf{v}_l$ is the latent factor vector of location l.

Intuitively, the above two latent factor vectors can be regarded as a low-dimensional representation for tourists and locations respectively, where the tourist vector $\mathbf{u}_u$ encodes the preferences information and the location vector $\mathbf{v}_l$ encodes the location property information. Once the two latent factor vectors are learned, we are able to infer tourist u's next visiting attractions by predicting the associated transport locations (i.e., metro stations or bus stops).

Besides the next tourist attraction prediction, another objective is to predict a tourist' immediate public transport boarding location, i.e., where the tourist wants to cease this

2. We assume the attractions are represented by the nearby metro stations or bus stops.

attraction visit. It can help tourists more efficiently organize their itinerary. In order to infer such an end point of a tourist attraction visit for more personalized services, we apply Markov model to capture the transition patterns among locations. We first define a pair of public transport alighting location l_x and the next boarding location l_y as a location transition $l_x \rightarrow l_y$. By summarizing location transition frequency, we build a location-location transition frequency matrix where each row represents an alighting location and each column represents a next boarding location. Following the way of representing tourists' preference for locations, we assign a binary variable $\varphi^s_{x,y}$ to represent the likelihood of transition from the alighting location l_x to the next boarding location l_y:

$$\varphi^s_{x,y} = \begin{cases} 1 & \text{if } r^s_{x,y} > 0 \\ 0 & \text{if } r^s_{x,y} = 0 \end{cases},$$

where $r^s_{x,y}$ indicates the total number of transitions from l_x to l_y. If the transition between l_x and l_y happened at least once, we can infer that l_y is likely to be the next boarding location of l_x. Intuitively, as the transition count grows, it is more likely that l_x and l_y are a pair of alighting and next boarding location. Following Eq. (9), we define the confidence of the likelihood that tourists leave at the location l_y given that they arrive at the location l_x as:

$$\theta^s_{x,y} = 1 + \alpha^s * r^s_{x,y}, \tag{11}$$

where α^s is a variable controlling the influence of the increase of the transition counts.

In order to infer tourists' preferences for alighting locations, as well as for alighting and the next boarding pairs (i.e., tours), we propose a model to co-factorize tourist-location visit matrix and location-location transition matrix, and accordingly learn latent factor vectors of tourists and locations. For tourists, we denote $\mathbf{u}_u \in \mathbb{R}^f$ as latent factor vector of tourist u, where f indicates the dimensionality of the latent factor vector. For locations, since we model alighting location and next boarding location separately, we assign a latent factor vector $\mathbf{v}_x \in \mathbb{R}^f$ and $\mathbf{v}'_y \in \mathbb{R}^f$ to each alighting location l_x and the next boarding location l_y respectively. So overall, we need to learn three sets of latent factor vectors for tourists, alighting locations and next boarding locations,[3] which share the same latent space to factor tourists' preferences and location transition likelihood.

To co-factorize both tourist-location visit matrix and location-location transition matrix, the loss function to be minimized is defined as:

$$\begin{aligned} J = &\sum_{u,x} \theta_{u,x}(\varphi_{u,x} - \mathbf{u}_u^\top \mathbf{v}_x)^2 + \lambda_1 \sum_{x,y} \theta^s_{x,y}(\varphi^s_{x,y} - \mathbf{v}_x^\top \mathbf{v}'_y)^2 \\ &+ \lambda_2 \Big(\sum_u \|\mathbf{u}_u\|^2 + \sum_x \|\mathbf{v}_x\|^2 + \sum_y \|\mathbf{v}'_y\|^2 \Big), \end{aligned} \tag{12}$$

where λ_1 is a variable controlling the effect of location-location transition matrix factorization, λ_2 is the regularization parameter that controls the complexity of the model.

3. Alighting locations and next boarding locations are from the same location set L, but play different roles in location transition.

4.2 Model Fitting

Due to the large size of the tourist-location visit count matrix and location-location transition matrix,[4] stochastic gradient descent (SGD), which is commonly used for conventional matrix factorization, cannot be directly applied. We thus adopt alternative least squares (ALS) [6] to efficiently fit the model. The basic idea is to first fix latent factor vectors of alighting locations and update those of tourists, and then alternate to update latent factor vectors of alighting locations by fixing those of tourists and next boarding locations. This procedure continues until the predefined number of iterations have been completed.

We denote latent factor matrix for tourists, public transport alighting locations and the next boarding locations by $\mathbf{U}_{|U|\times f}$, $\mathbf{V}_{|L|\times f}$ and $\mathbf{V}'_{|L|\times f}$, where each row of the matrices corresponds to the latent factor vector of a tourist, an alighting location, and a next boarding location respectively. $\mathbf{\Theta}_u$ denotes a $|L| \times |L|$ diagonal matrix where the values on the diagonal represent the confidence of tourist u's preferences for the corresponding alighting locations (see Eq. (9)). $\mathbf{\Phi}_u$ represents tourist u's real binary preferences for alighting locations. To update tourists' latent factor vectors, we compute the gradient of J with respect to each tourist u's latent factor vector, and then obtain the updated latent representation:

$$\mathbf{u}_u = (\mathbf{V}^\top \mathbf{\Theta}_u \mathbf{V} + \lambda_2 I)^{-1} \mathbf{V}^\top \mathbf{\Theta}_u \mathbf{\Phi}_u, \tag{13}$$

where I is identity matrix. In the same way, the updated latent factor vector of an alighting location l_x is computed as follows:

$$\begin{aligned} \mathbf{v}_x = (\mathbf{U}^\top \mathbf{\Theta}_x \mathbf{U} + \lambda_1 \mathbf{V}'^\top \mathbf{\Theta}^s_x \mathbf{V}' + \lambda_2 I)^{-1} (\mathbf{U}^\top \mathbf{\Theta}_x \mathbf{\Phi}_x \\ + \lambda_1 \mathbf{V}'^\top \mathbf{\Theta}^s_x \mathbf{\Phi}^s_x), \end{aligned} \tag{14}$$

where $\mathbf{\Theta}_x$ and $\mathbf{\Theta}^s_x$ are $|U| \times |U|$ and $|L| \times |L|$ diagonal matrices where the values on the diagonal represent the confidence of the corresponding tourists' preference for this alighting location l_x and the likelihood of the corresponding boarding location of l_x (see Eq. (11)) respectively. $\mathbf{\Phi}_x$ and $\mathbf{\Phi}^s_x$ are vectors containing tourists' binary preference for l_x and l_x's binary transition variable to boarding locations respectively. Similarly, the latent factor vectors of alighting locations are updated as follows:

$$\mathbf{v}'_y = (\mathbf{V}^\top \mathbf{\Theta}^s_y \mathbf{V} + \frac{\lambda_2}{\lambda_1} I)^{-1} \mathbf{V}^\top \mathbf{\Theta}^s_y \mathbf{\Phi}^s_y, \tag{15}$$

where $\mathbf{\Theta}^s_y$ is $|L| \times |L|$ diagonal matrix where the values on the diagonal are the confidence of the likelihood that l_y is the next boarding location of the corresponding alighting locations. $\mathbf{\Phi}^s_y$ is a vector with alighting locations' binary transition variable to l_y.

After the above latent factor vectors of tourists, alighting locations and the next boarding locations are learned, we are able to infer tourists' next visited attractions by predicting the corresponding alighting locations. Furthermore, it is feasible to infer where tourists leave the attractions by predicting the corresponding next boarding locations. Such

4. Optimization should consider all tourist-location and location-location pairs, including both observed data and unobserved data, i.e., the missing values in the visit count matrix.

location-based tour predictions provide rich opportunities to significantly enhance tourists' experience, e.g., by suggesting top-N interesting attractions or pushing location-aware promotion information.

Note that the matrix inversion used in the ALS algorithm is an expensive operation, and its time complexity is normally assumed $\mathcal{O}(f^3)$ [6], where f is the dimensionality of the latent factor vector. Accordingly, the overall time complexity of one iteration is $\mathcal{O}((|U|+|L|)f^3+|U||L|f^2+|L|^2f^2)$, where $|U|$ and $|L|$ are the number of identified tourists and the number of locations (i.e., subway stations and bus stops) respectively. To reduce the above complexity and speed up the ALS by eliminating the $\mathcal{O}(f^3)$ term, the fast matrix factorization techniques [7] can be adopted in the practical implementation.

4.3 Location Filtering

The location set L in the proposed analytics model is crucial, which is mainly obtained from the tourist' transport riding traces. However, not all the locations are the real destinations (i.e., tourist POIs) in their travel plans. For example, some locations are the places where they stay (e.g., hotels or other temporary resting sites), some are simply the stopover for bus or subway transit, and some locations are even the places tourists wrongly reach.

In practice, the data-driven approach can be used to conduct the location filtering task, and the following filtering strategies can be selectively used:

- Accommodation places: By common sense, a tourist normally starts a day tour from his or her staying place, and come back after finishing the day tour. Hence, the start and the end station of a tourist can be excluded from the location set L.
- Transit or wrong places: If a tourist gets on a train or bus very shortly after he or she gets off from the same or a very close location, it can be considered as a transit behavior or wrongly reaching a place. Accordingly, given a tourist's boarding and last alighting location are too close and meanwhile the duration are too short, the corresponding locations can be excluded from the location set L. In practice, a temporal threshold and a spatial threshold can be set.
- Ambiguous places: Some records may show that a tourist alights from one station, and after a significant time period, he or she either boards or alights at another station that is far away from the previous one. Such stations can be excluded from the location set L, as it is hard to determine the exact places the tourist visited.

Note that the above described filtering strategies may wrongly discard some useful locations for understanding tourist' travel patterns or behaviors, but they can directly help to provide more reliable data for the preference analytics. After the location filtering, we can simply segment each tourist's riding trajectory into tours. Since our work uses public transport data, each tour is described by subway or bus stations. While each station may indicate one or several attractions that are geographically close to each other, the analytics and recommendation application can show all such information. In addition, the public transportation data are usually maintained in Hadoop systems due to their large size, and all the above operations on the raw data may need to be implemented using MapReduce or similar programming models.

In short, by leveraging on the identified tourists' records from the public transport data, a tourist recommendation model is constructed for understanding and predicting the tourist preferences. We propose the model and its ALS-based fitting techniques to co-factorizes both tourist-location visit matrix and location-location transition matrix, where the location filtering strategies are designed to better support the tourist preference analytics.

5 Empirical Experiment

We have conducted the comprehensive experiments to evaluate the performance of the proposed framework, where we take Singapore as an exemplary case and mainly use the data from its public transport system. In this section, we first give an overview of Singapore's public transport system and its transport data, and then present the experiment results for tourist identification and preference analytics using the corresponding data.

5.1 Singapore Public Transport System and Dataset

In Singapore, the public transport system mainly consists of the subway service and the bus service. The deployed automatic ticketing system for the city's public transportation uses a contactless ticketing card, called EZ-Link card, to charge the trip fares at all subway stations and bus stops. Such a ticketing system naturally tracks each commuter's riding, and fare is dynamically calculated based on the total travel distance. As EZ-Link card users can enjoy a fare discount, nearly all Singaporean residents use the card to take subway and bus. For the subway service, the ticketing card is required to tap in and tap out at the gantry to calculate the current trip fare. EZ-Link card is also a good choice for tourists, especially for those who stay for a few days and frequently travel in the city. Buying an EZ-Link card requires a minimum payment of 12 Singapore dollars including a non-refundable 5 Singapore dollars of issuing cost. A commuter, especially the short-stay tourist, may opt to purchase a one-time ticket with cash and use in the same way as an EZ-Link card. The main difference is that the per-ride price of using one-time ticket is higher than using the normal EZ-Link card. In addition, there is another type of EZ-Link card, called concession card, which is only eligible for the Singapore citizens and permanent residents, and it offers a larger discount on trip fares specifically for local students and senior citizens.

All the transaction records for all the EZ-Link cards and one-time tickets can be automatically collected by the backend public transport system. The dataset contains records from both bus and subway rides that are paid with regular EZ-Link cards, concession EZ-Link cards and one-time tickets. Each record contains multiple fields, and for this work, we only use several selected fields that are summarized in Table 1. Some explanations on the fields of *Transport Mode* and *Payment Mode* are given below:

TABLE 1
Dataset Schema

Field	Description
Card_Number_E	Encrypted Card ID
Transport_Mode	BUS or SUBWAY
Entry_Date	Date when ride started
Entry_Time	Time when ride started
Exit_Date	Date when ride ended
Exit_Time	Time when ride ended
Payment_Mode	Method of payment
Origin_Location_ID	Starting location of the ride
Destination_Location_ID	Ending location of the ride

- *Transport_Mode*: Its possible values include either SUBWAY or BUS, and the corresponding Origin_Location_ID and Destination_Location_ID are subway station and bus stop respectively. All the subway stations and bus stops can be directly mapping to the geographical location (i.e., in terms of latitude and longitude) using the published official data.
- *Payment_Mode*: Its possible values include CSC, PASS or STANDARD. CSC refers to normal adult EZ-Link cards, PASS refers to concession cards, and STANDARD refers to one-time tickets.

In the experiment, we mainly utilize the above described data and information to conduct the experiments for tourist identification and preference analytics.

5.2 Tourist Identification

5.2.1 Data Preparation

The entire experiment is conducted using three months' riding records from 5.1 million distinct commuters and their 462 million trips. Compared to around 5.3 million local residents in Singapore, it shows a good coverage of the public transport population in Singapore. We first preprocess the dataset by excluding the commuters with less than 6 tuples of travelling records with EZ-Link card (i.e., the payment mode is CSC), as normally tourists with a few times of public transport riding would prefer using one-time tickets (i.e., the payment mode is STANDARD) to saving the issuing cost. After the pre-processing step, the leftover records contains 1.7 million commuters with a total of 49.5 million transactions. To obtain a small group of the labeled dataset for the tourist identification task, we asked the local people who know Singapore well to manually label 1000 tourists using their riding records, where the criteria include the number of active days, daily travel routes (e.g., the detailed origin and destination stations) and the staying periods at each station. The corresponding Kappa value is 0.92. Moreover, we also randomly sampled around 250 thousand local commuters from the total of 420 thousand concession card users, as only the local people (typically including Singapore citizens and permanent residents) can purchase concession card.

5.2.2 Parameter Estimation

We first define a key parameter ρ to describe how frequent a local commuter turns to use one-time ticket due to different personal reasons, the parameter ρ quantifies the ratio between the number of one-time ticket records and the number of regular records made by local commuters. The basic idea to estimate ρ is based on the key observation that tourists are very unlikely to visit stations that are located in residential areas with few shopping, hotel and restaurant facilities. We thus first ask the local people to specify a number of such stations. At these stations, the dataset shows that there are still one-time ticket records showing these stations as destinations, which convinces us that such rides are mainly made from local commuters rather than tourists. Table 2 summarizes the count of one-time ticket records n_i^s, the count of regular EZ-Link ticket records n_i^r, and their ratio at such stations.

TABLE 2
Statistics of Non-Tourist Stations

Station Name	n_i^s	n_i^r	$\frac{n_i^s}{n_i^r}$
Marymount	6218	629435	0.009879
Yio Chu Kang	20361	2067636	0.009847
Cove	1817	189873	0.009570
Buangkok	7454	787463	0.009466
Layar	345	37211	0.00927
Oasis	489	53696	0.009107
Labrador Park	2473	292858	0.008444
Tongkang	1295	158299	0.008181
Compassvale	2705	358175	0.007552
Dover	8963	1247247	0.007186

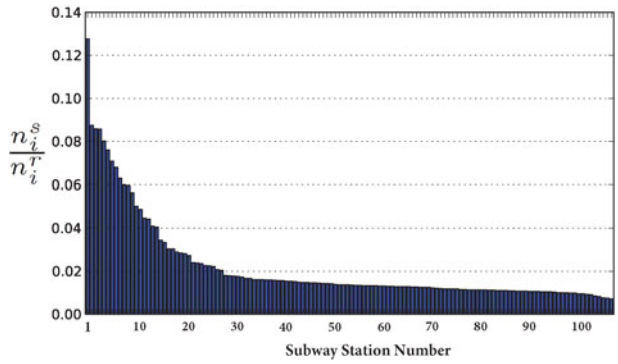


Fig. 3. Ratio distribution at stations in a descending order.

From Table 2, we see that the ratio $\frac{n_i^s}{n_i^r}$ is relatively stable at these stations, which partially verifies our assumption that the parameter ρ is stable and independent with stations. Hence, we can make a fair estimate on ρ using such information. Moreover, we see that the ratio is still increasing from the bottom station to the top ones, it is probably because more tourists use one-time tickets at the top stations in the table. Fig. 3 further shows the ratio distribution in a descending order for all the subway stations in Singapore. We see that the distribution exhibits a long-tail pattern, where the ratio at the tail are getting closer to the actual parameter ρ. Accordingly, we use the ratio at the last station, called *Dover* station, as an estimate of parameter ρ. *Dover* station is located next to Singapore polytechnic school, and its nearby area does not have any famous attractions or tourism facilities but only residential buildings. Hence, it is reasonable to assume that most of the one-time ticket users are local commuters rather than tourists. As we see from Table 2, the ratio at Dover station is 0.007186, and we use it as an estimate of ρ in the experiment.

TABLE 3
Ranked Stations Based on Initialized Station Scores

Station Name	Initialized Score
Changi Airport	0.213668
Marina Bay	0.145012
Clarke Quay	0.144702
Bayfront	0.128008
Little India	0.118879
Chinatown	0.113837
HarbourFront	0.106443
Bras Basah	0.104787
Esplanade	0.099637
Orchard	0.098623
Lavender	0.093104
Farrer Park	0.081844
Promenade	0.079080
Bugis	0.070973
City Hall	0.064815

Accordingly, the number of tourists using one-time tickets at the station ith station can be computed as follow:

$$n_i^t = n_i^s - n_i^r \cdot \rho, \tag{16}$$

where n_i^s is the number of one-time ticket records and n_i^r is the number of regular ticket records at the ith station.

5.2.3 Station Score Initialization

After determining the parameter ρ and n_i^t, we further compute the initial scores for all the stations using Eq. (1). Table 3 summarizes the top 15 stations in terms of the decreasing order of the computed initial scores. All of the 15 stations are located nearby the famous attractions and POIs in Singapore. The top ranked station is Changi Airport, which is the only international airport of the city. It is a reasonable result, as most tourists start or complete their trips in this airport whereas local commuters do not often visit it except they travelling overseas. Moreover, Table 3 also shows that some must-visit venues, such as Orchard station (the most famous shopping mall area) and City Hall station (city's central area) are ranked not that high, as local Singapore citizens also frequently visit such places for shopping or business. In short, the initialized scores for the stations can be used to well capture the characteristics of the station and identify tourists from local commuters.

5.2.4 Tourist Identification

We run the designed iterative propagation learning algorithm on the proposed SCR graph with the initialized station scores. As the configurable parameters for controlling the updating rate, the parameter α is usually set lower than the other two parameters β and γ, as it is used to directly update the probability distribution for the unlabeled commuters. In our experiment, α is set to 0.7, and the other two are set to 0.9 by using a trail-and-error approach. The iteration number is set to 120. Figs. 4a, 4b, 4c and 4d showcase the four typical stages of the propagation process respectively, namely initial stage, commuter updating stage, station updating stage and final stage. The results from the final stage will be used to classify the commuters.

To fully evaluate the classification performance, we adopt both micro-averaged F1 and macro-averaged F1 measures [8] as the main metrics. Meanwhile, we compare the proposed algorithm with the two classification methods below:

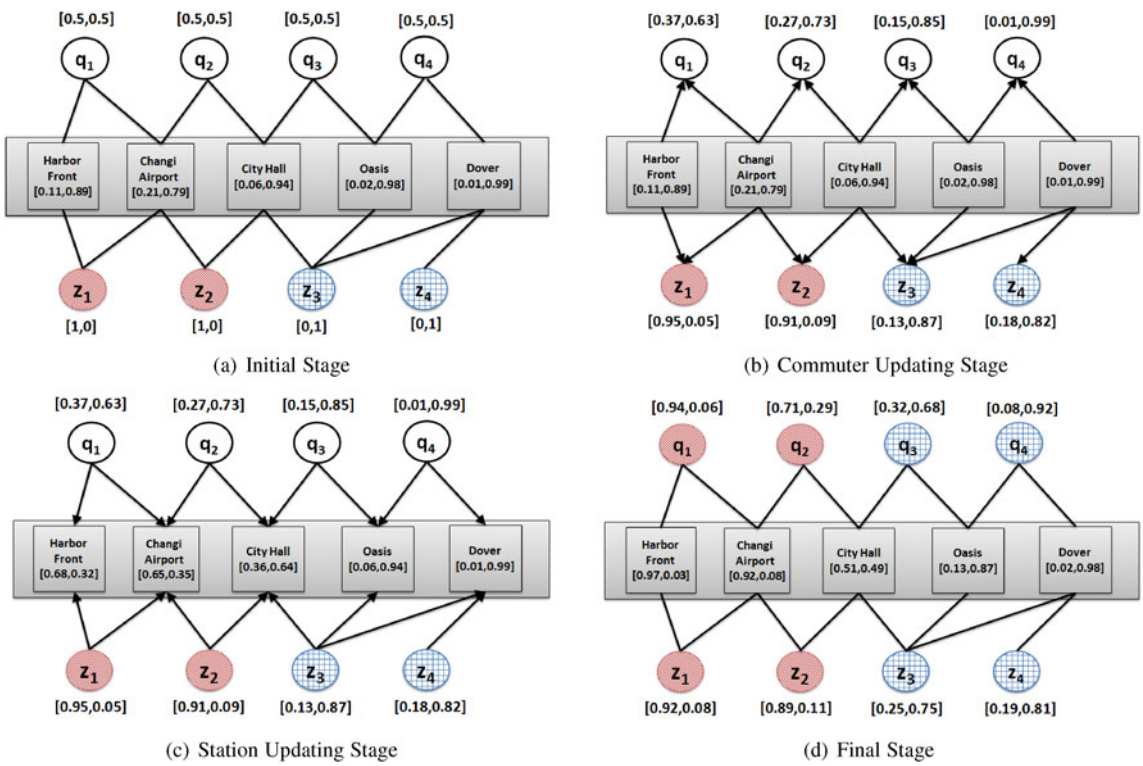


Fig. 4. Exemplary propagation process with four typical stages.

TABLE 4
Performance Comparison on the Tourist Classification

	SVM		FTF		Ours	
p%	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1
5%	0.57984	0.8415	0.6109	0.8419	**0.6267**	**0.8504**
10%	0.5917	0.8420	0.6263	0.8464	**0.6572**	**0.8538**
15%	0.6144	0.8411	0.6441	0.8433	**0.6677**	**0.8560**
20%	0.6199	0.8480	0.6758	0.8504	**0.6962**	**0.8575**
25%	0.6286	0.8402	0.6956	0.8459	**0.7154**	**0.8549**

- Fitting The Fits (FTF) [9] model: a well-known iterative inference algorithm uses labeled data and unlabeled data, which assigns predicted values to the observations whose responses are missing (unlabeled data) in each iteration, and then incorporates the predictions appropriately in the subsequent steps.
- Support Vector Machine (SVM) model [10]: a well-known supervised classification algorithm, and the labeled 1 thousand tourists data are used as the positive class to train the model.

Table 4 gives the performance of the designed algorithm against another two methods, where the parameter p represents the percentage of data used as training data. For example, when p equals 5 percent, it means 5 percent of labeled data is used to train the model and 95 percent is used to validate the model. Table 4 shows that our algorithm achieves 0.8549 and 0.7154 on micro-averaged F1 and macro-averaged F1 measures respectively, when 25 percent of labeled data is used for the model training. Furthermore, for all the p values, our algorithm achieves the higher scores on both F1 measures compared to another two methods. Specifically, SVM exhibits the worst performance, which is probably caused by the limited labeled data available. The designed algorithm achieves the best performance, as it iteratively propagates the labeled information and the prior knowledge (e.g., station scores) to unlabeled dataset, and recursively makes use of such data in the next step of iteration. Lastly, we see that for all the three algorithms, the more training data used, the better performance can be achieved.

In short, the experiment results show that the designed module performs well for the tourist identification task using the public transport data, where the SCR graph and the iterative propagation learning algorithm achieve their design objectives. Finally, a total of 206 thousand tourists together with their travel trajectories are identified from the three months' public transport data, which are used to conduct their preference analytics.

5.3 Tourist Preference Analytics

5.3.1 Data Preparation

The travel records of each identified tourist are first segmented into individual tours, represented by the pairs of alighting location and the immediate boarding location. To evaluate the performance of the proposed model, we used every tourist's first 70 percent tours (in chronological order) as training data for model construction, and the rest are for testing. Moreover, we do not consider the most popular tourist attractions, as these must-sees are visited by almost every tourists and naturally appear in any tourist location recommendation results. This experiment investigates the effectiveness of the personalized preference analytics, and thus considering such extremely popular attractions make the personalization weaker. Moreover, it is a more interesting and challenging task to discover the locations not generally popular but favored by specific tourists. Hence, in our final evaluation results, we sort all locations based on the number of tourist visits and remove the top 20 most popular ones.[5] The experiment results thus also show how the proposed model copes with this challenge.

5.3.2 Baselines

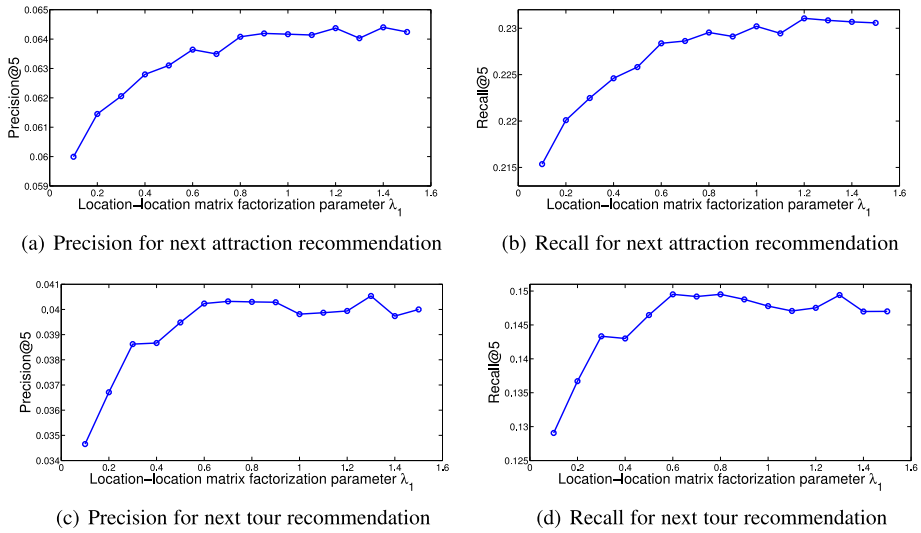
We compare our model with four popular baselines summarized as follows:

- Popularity based model (POP) [11]: This method predicts and recommends tourists' next visiting attraction and tour (i.e., alighting location and next boarding location pair in our definition) based on the corresponding popularity, e.g., the number of visits. It means the most popular locations and tours are recommended to all tourists without any personalization.
- Markov model (Markov) [12]: This method makes recommendation based on Markov property, i.e., the probability that a tourist visits a location x based on her last visited location y. Specifically, two location transition probability matrices are built, where one matrix records the probability from one alighting location to another alighting location (for next attraction recommendation), and the other matrix records the probability from alighting location to the next boarding location (for entire tour recommendation).
- User-based collaborative filtering (UCF): This is the traditional personalized recommendation method that predicts users' preference for items using similar users' information, where user-user similarity is calculated based on their historical visit behaviors. Two UCF models are built for next attraction recommendation and next tour recommendation respectively.
- Graph-based POI embedding model (GE) [13]: This is a graph-based embedding model that jointly captures the sequential effect, geographical influence, temporal cyclic effect and semantic effect by embedding the corresponding relational graphs.

5.3.3 Metrics

The proposed model is able to recommend the top-N locations and location pairs that are not visited by a tourist but are most likely visited in his or her immediate plan. To measure the accuracy of recommendations, we used two metrics: precision@N, which is the ratio of the successfully predicted locations (or location pairs) to the top-N

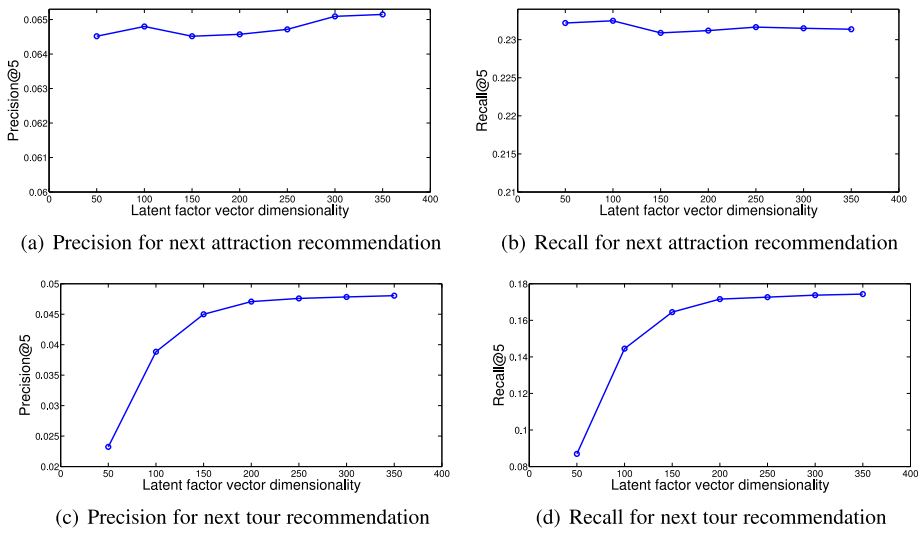
5. The locations are represented by metro stations or bus stops. The top 20 locations correspond to less than 10 must-see POIs in Singapore, verified with the background knowledge.



(a) Precision for next attraction recommendation (b) Recall for next attraction recommendation

(c) Precision for next tour recommendation (d) Recall for next tour recommendation

Fig. 5. Performance with varying location-location matrix factorization parameter λ_1 (Top-5 recommendation).



(a) Precision for next attraction recommendation (b) Recall for next attraction recommendation

(c) Precision for next tour recommendation (d) Recall for next tour recommendation

Fig. 6. Performance with varying latent factor vector dimensionality (Top-5 recommendation).

recommendations, and recall@N, which defines the ratio of successfully predicted locations (or location pairs) to the number of locations (or location pairs) to be predicted. We will demonstrate how the performance of models varies with different values of N.

5.3.4 Parameter Tuning

We refer to the next attraction (public transport alighting location) and next tour (alighting location and next boarding location pair) recommendation as step 1 and step 2 recommendation respectively. We first study how location-location matrix factorization (with varying λ_1) influences the performance of recommendation. We set latent factor vector dimensionality, regularization parameter, confidence parameters α^u and α^s to 100, 0.04, 1 and 1 respectively. When top-5 recommendations were provided, Fig. 5 shows precision and recall for step 1 and step 2 recommendations with λ_1 varying from 0.1 to 1.5 (0.1 as the increment). We observe that when λ_1 becomes larger, the precision and recall for both steps increase. This is because more weight for the location-location matrix factorization improves the quality of latent factors of next boarding locations, which in turn better fits latent factors of alighting locations and tourists. However, when λ_1 reaches a certain threshold, the precision and recall becomes stable and larger λ_1 values do not bring in evident improvements. It is worth noting that when λ_1 further increases, the performance starts slightly declining. Specifically, for step 1 recommendation, we notice that from $\lambda_1 = 0.8$, we get the stable results, while for step 2 recommendation, from $\lambda_1 = 0.6$ is a good choice. In the following experiments, we choose the optimal λ for our model.

Next, we study the influence of latent factor vector dimensionality f when top-5 recommendations are provided (see Fig. 6). Other parameters like location-location matrix factorization parameter λ_1, regularization parameter λ_2, confidence parameter α^u and α^s are set to 0.9, 0.04, 1 and 1, respectively. For step 1 (next attraction) recommendation, we notice that the performance of our model is insensitive to the latent factor vector dimensionality. We also test the extreme values and observe that when the dimensionality is smaller than 10, the precision and recall start evidently

TABLE 5
Results for Next Attraction Recommendation

	Top-5		Top-10		Top-15	
	Precision	Recall	Precision	Recall	Precision	Recall
POP	0.025	0.096	0.022	0.173	0.021	0.247
Markov	0.03	0.116	0.025	0.182	0.021	0.258
UCF	0.036	0.142	0.030	0.210	0.024	0.260
GE	0.051	0.196	0.059	0.231	0.033	0.271
Ours	**0.063**	**0.226**	**0.076**	**0.272**	**0.042**	**0.302**

TABLE 6
Results for Next Tour Recommendation

	Top-5		Top-10		Top-15	
	Precision	Recall	Precision	Recall	Precision	Recall
POP	0.015	0.060	0.014	0.108	0.012	0.142
Markov	0.019	0.069	0.014	0.126	0.013	0.152
UCF	0.024	0.093	0.016	0.146	0.015	0.169
GE	0.035	0.133	0.020	0.186	0.023	0.211
Ours	**0.048**	**0.178**	**0.029**	**0.221**	**0.027**	**0.257**

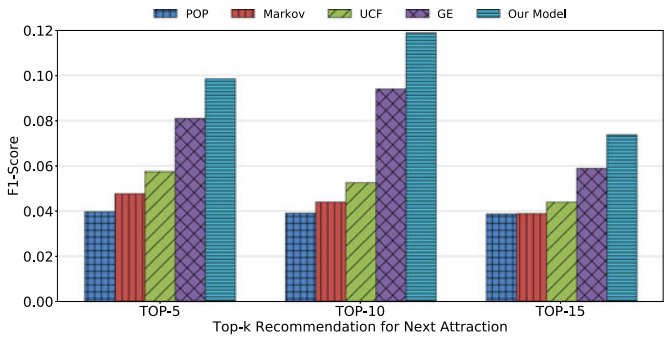
decreasing. On the other hand, step 2 (next tour) recommendation is more sensitive to the dimensionality. From Fig. 6c and 6d we notice that the performance is improved significantly with the increasing latent factor vector dimensionality. For instance, the precision and recall get increased by 103.02 and 97.47 percent respectively when the dimensionality changes from 50 to 200.

In short, if the system is only interested in next attraction recommendation, a small latent factor vector dimensionality like 50 to 100 may be chosen so that the model can be trained fast; otherwise, the optimal dimensionality for the tour recommendation task is 200, from which the the performance become stable and the higher dimensionality does not improve it evidently but would incur a higher computational overhead.

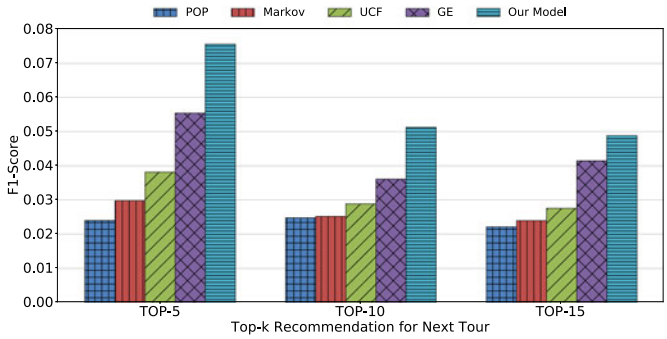
5.3.5 Comparison Results

We compare our model with the four baselines, i.e., POP, Markov, UCF and GE, where the optimal parameters presented in previous subsections are used for model training. Table 5 summarizes the comparison results for next attraction recommendation when different recommendation list is provided. As mentioned, the very top destinations that will be visited by nearly all tourists (e.g., city hall) are omitted in the performance measurement, so that the power to predict more "personalized" locations by different algorithms can be well reflected. As expected, POP performs worst. This is because tourists' visit behavior is heterogeneous, therefore, simply recommending the most popular attractions does not necessarily meet tourists' real needs. For instance, it is observed that the most popular attractions in Singapore aggregate at the city's central area, but some tourists prefer visiting natural landscape such as botanic gardens or Bukit Timah nature reserve. The popularity based approach fails to capture such personalized cases.

Due to the same reason, the performance of Markov approach is not promising either, although by learning transition patterns of attractions, on average, Markov improves POP by 11.23 and 10.08 percent in terms of precision and recall respectively. It is worth noting that one issue with Markov is the data sparsity, i.e., due to limited data records, the transition probabilities cannot be derived with a high confidence. GE takes into account the spatiotemporal context information and alleviates the data sparsity issue. It evidently improves Markov, POP and UCF, while it does not perform better than our model. One of the possible reasons is that the moving behaviors and patterns of tourists, especially the foreign tourists in the city like Singapore, may not strongly exhibit the desired temporal cyclic effect and sequential patterns.



(a) F-score for next attraction prediction



(b) F-score for next tour prediction

Fig. 7. F-score comparison for top-k predictions.

In all cases, our approach outperforms the baselines due to two key designs: (1) Our model is applied to learn tourists' preference for personalization by handling the implicit feedback information in transport data for top-N recommendation, which is naturally suitable for our tasks. (2) The alighting locations and the next boarding locations are treated separately, and the location transition matrix is factorized (jointly with tourist-location matrix factorization) to learn tourists' location preference. Overall, our approach improves the baselines by at least 23.53 and 11.44 in terms of precision and recall.

We also summarize the precision and recall for next tour recommendation in Table 6. Similar trends are observed: GE performs better than Markov and UCF, which outperform the weakest baseline POP; our approach outperforms the four baselines. To balance the precision and recall for an overall performance measurement, Fig. 7 further shows the F-score comparison among the four approaches for both next attraction and next tour predictions.

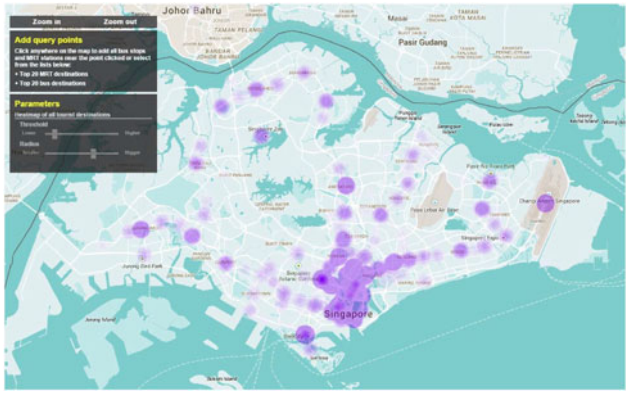


Fig. 8. User interface overview.

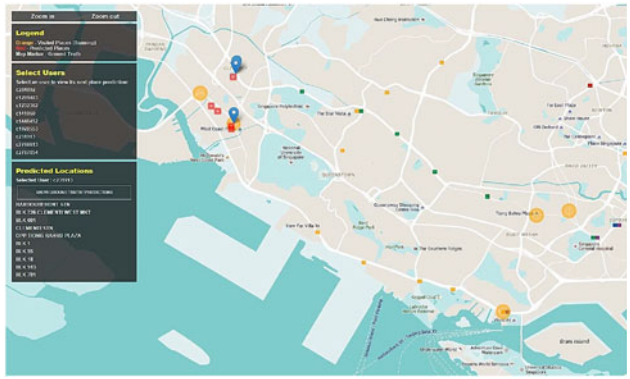


Fig. 9. Tourist preference analytics and prediction.

5.4 User Interface

An informative user interface is developed to support different stakeholders to access the spatial and temporal characteristics of the identified tourist information, which provides users an interactive and intuitive way to understand the analytics results. Following our previous design for tourist tracking [14], Fig. 8 shows how users can access the analytics results from the perspective of destinations, where a heat map is superimposed on Singapore map to visualize the distribution of the identified tourists' destinations in the city. For example, the heat map shows that a couple of circles found in north part of the island, where almost no local attraction or shopping mall nearby. We then found that it is mainly because the places provide direct bus services between Singapore and Malaysia, where many tourists board or alight there. Moreover, users can click any place on the map to query the detailed tourist information surrounding that location (e.g., tourist lists and their staying periods).

Fig. 9 shows how users can access the preference analytics results from the perspective of individual tourists. On the left side panel, a list of tourists is shown, where all the tourist information are anonymized in the current system. Once clicking on a tourist, the places he or she has visited will be shown on the map (orange circles), and meanwhile the recommended locations for this tourist will be presented as well using the proposed preference analytics model. For the demonstration purpose, Fig. 9 illustrates both the recommended locations (red squares) and the tourist next-visit places (blue landmarks), where we can see two locations are correctly recommended. Note that in the running system, only the visited locations (orange circles) and recommended locations (red squares) can be shown on the user interface in real time.

6 Discussion

6.1 Possible Limitations

The algorithms presented in this paper are mainly designed for specifically identifying and analyzing tourists using the public transport data, which may need to be further extended to handle more general cases. For example, some heuristics and hard conditions are imposed to identify the tourists with a high confidence, which inevitably fail to recognize some actual tourists. Moreover, for the preference analytics, especially when defining and constructing the tourist-location visit matrix, we only consider the count of visits between tourists and stations but omit some useful information, such as their staying periods and sequence of visiting stations. Such information can be an effective indicator to show how much a tourist favors different locations by comparing it with other tourists. In a similar vein, to increase the accuracy of the tourist identification, the decision of labeling tourists can be made by considering more about their staying period and visiting sequence information. Accordingly, more accurate models for tourist identification and preference analytics might be able to build and reveal more interesting "tourism" patterns. Besides, the current tourist identification solution is only applicable to the transport data, while how to concurrently using other types of data, such as GPS and social network data, can be an interesting question and worth to be investigated.

On the other hand, some configurable parameters used in the current algorithms, such as the confidence parameters α^u and α^s used in the preference analytics model, are important to the system. Such parameters currently are determined mainly using trail-and-error approach. Some machine learning techniques can be introduced to automatically derive these parameters based on the historical measurements and accordingly provide a better robustness. For example, the cost-sensitive classification models can be employed to reflect the application-specific tradeoff between false-negative and false-positive errors. In addition, the current design is mainly based on the transport system having the tap-in/tap-out control, while for the systems with the tap-in control only, some predictive models may need to be designed and deployed to infer passenger alighting location and traveling time.

6.2 Transportation Data Enrichment

In this work, we only use the public transportation data from the subway and bus system. The richer and new modes of transportation data can be introduced to further improve the system performance. For example, taxi data can be considered to jointly analyze the tourists, as many foreign tourists may choose taxi as their personal transport means in the city. Currently, it is relatively hard to continuously track an individual tourist's traces on different taxis and link taxi data with the public transport data. However, the emerging new taxi payment methods may help to tackle these issues and facilitate the taxi data collection process: for example, most

taxis in Singapore recently start supporting passenger to pay the fare using the near-field-communication (NFC) enabled cards and EZ-Link cards.

6.3 Real-Time Analytics and Implementations

The empirical results presented in this work are mainly based on the offline analytics, i.e., using historical transport data to conduct tourist identification and preference analytics. While such offline analytics and experiments are adequate for validating the proposed framework and demonstrating how the public transport data can be used for tourist analytics, additional enhancements and models may be needed to support real-time analytics tasks, whose time cost and efficiency need further careful evaluation. For example, the tourist identification system may need to be slightly modified to support early tourist detection that uses the first several trips of tourist data. Similarly, tourist preference analytics system may need to utilize the latest tourist information to make faster recommendations. Such system enhancement issues provide important directions for future work.

6.4 Potential Applications

The prompt identification of tourists and their preference can support a number of potential applications and benefit different stakeholders, such as tourists and tourism administrators. For example, accurate and real-time recommendation functions can be integrated into travel-planning systems, where tourists can use the information to make better travel decisions (e.g., deciding where to go and visit for the next days). An accurate and fine-grained understanding of tourist personal preference can also help tourism service providers to engage better with tourists. For example, the service providers can design the personalized travelling packages, and push such product advertisements on the screen of the gantry of the stations or inside the carriages. Using the aggregated tourist travelling statistics, the relevant government agencies can intelligently deploy new bus fleets or walking routes specifically designed for tourists, thereby enhance tourist experiences by providing more convenient and comfortable travelling solutions.

7 Related Work

7.1 Tourism Research and Tourist Data

Due to the complex nature and characteristics of tourism [15], many research efforts require empirical data from tourists to analyze and model the spatio-temporal behavior of tourists [16]. Traditionally, the tourist data were mainly collected using labor-intensive and error-prone methods, including direct observation [17], personal interviews [18] and space-time diaries [19]. Recently, researchers have utilized the data from global navigation satellite systems (e.g., GPS data) [20], cellular systems (e.g., call detail record data) [21], social media (e.g., geotagged images in Flickr) [1], and sensor networks (e.g., bluetooth data) [4]. However, their data collection systems mainly adopt the participatory sensing strategy, which suffers the risk of self-selection bias (e.g., certain groups of tourists may show a low degree of presence in the collected data). In other words, the existing data collection methods are hard to scale up and difficult to cover the majority of tourists. Following our previous work [22] on this topic, we innovate in utilizing public transportation data to collect information and conduct analytics on tourists, which provides a non-intrusive and inexpensive solution to tackle this issue.

7.2 Tourist Preference Analytics and Recommendation

The tourist preference analytics and recommendation has been a popular and lucrative topic due to its economic importance, where the collaborative filtering techniques and the POI recommendation methods have been well studied [23], [24], [25], [26]. For example, Yin et al. propose a latent class probabilistic generative model to understand user interests and preference using the "check-in" data from location-based social networks [27]. Besides the models that utilize geo-social information [28], [29], temporal information [30], [31], semantic information [32], [33], and deep learning models have been utilized to improve POI recommendation performance [34], [35]. For example, the model SH-CDL [35] performs deep representation learning for POI recommendation, where heterogeneous features of the POIs and the collective preferences of the public in the target regions are well exploited.

Different from most of the existing POI recommendation studies that mainly leverage on user check-in activities and social influences information, our tourist preference analytics model and the entire framework design target on utilizing the information solely and directly derived from the public transport system rather than relying on any other information systems or sources. However, it possibly improves the current model performance by introducing and combining more information from both user side and location side, such as the social influences from location-based social networks (LBSNs) or the explicit rating score on the POIs, together with some POI recommendation algorithms.

7.3 Transportation Data Analytics

Taking advantage of electronic ticketing systems and on-vehicle telematics, the large amount of public transportation data in many cities are collected automatically and become publicly available, typically including smart card data [36] and taxi data [37]. By leveraging on the transportation data, researchers primarily focus on understanding and quantifying commuters' moving patterns for the improvement of city operations. For example, smart card data are used to study commuter travel patterns [38], passenger segmentation [39], passenger route estimation [40], and subway boarding analysis [41]. Similarly, taxi data are used to study urban planning [42], passenger wait time prediction [43], taxi and passenger queuing [45], taxi trip clustering [44] and driver recommender systems [46]. However, a few studies focus on analyzing transportation data to understand the characteristics of a special group of commuters. To our best knowledge, no previous work has conducted the tourist identification and preference analytics using the public transport data.

8 Conclusion

In this paper, we have introduced *TourSense* framework that first identifies tourists and subsequently conducts their preference analytics using city-scale public transportation data. The SCR graph together with the iterative propagation learning is proposed to effectively recognize tourists from public commuters. After that, a tourist preference analytics model is constructed to predict next attraction and tour. We have shown the promise of this approach via using the city-scale data from Singapore public transportation system. In the experimental results, the Macro and Micro F1 scores of the proposed tourist identification approach achieves 0.8549 and 0.7154 respectively, and meanwhile the proposed preference analytics model improves the baselines in terms of both precision and recall. An interactive and informative user interface is developed to help access and visualize all the analytics results.

On a broader canvas, the proposed framework demonstrates the feasibility of recognizing and analyzing different groups of public commuters, such as tourists, business travellers, local citizens, or even foreign workers. We believe that many other insights of practical interest (e.g., the different travel demands and behaviors between tourists and business travellers) can be investigated using the proposed framework and the public transport data. Moreover, this work reveals many unique advantages of transport data over other information sources (e.g., social media data), typically including a good coverage of population, timeliness of information, and the usefulness of the transportation infrastructures (e.g., subway gantries or bus stops can be potentially used to distribute the analytics results).

Acknowledgments

This research is partially supported by the Humanities and Social Sciences Foundation of the Ministry of Education of China (No. 17YJCZH116), the National Natural Science Foundation of China (No.61702039, 61807003), and the Fundamental Research Funds for the Central Universities.

References

[1] I. Cenamor, T. de la Rosa, S. Núñez, and D. Borrajo, "Planning for tourism routes using social networks," *Expert Syst. Appl.*, vol. 69, pp. 1–9, 2017.

[2] A. Chua, L. Servillo, E. Marcheggiani, and A. V. Moere, "Mapping cilento: Using geotagged social media data to characterize tourist flows in southern italy," *Tourism Manage.*, vol. 57, pp. 295–310, 2016.

[3] G. Kim and L. Sigal, "Discovering collective narratives of theme parks from large collections of visitors' photo streams," in *Proc. ACM SIGKDD Conf.*, 2015, pp. 1899–1908.

[4] M. Versichele, et al., "Pattern mining in tourist attraction visits through association rule learning on bluetooth tracking data: A case study of ghent, belgium," *Tourism Manage.*, vol. 44, pp. 67–81, 2014.

[5] J. Steenbruggen, E. Tranos, and P. Nijkamp, "Data from mobile phone operators: A tool for smarter cities?" *Telecommun. Policy*, vol. 39, no. 3/4, pp. 335–346, 2015.

[6] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proc. IEEE Int. Conf. Data Mining*, 2008, pp. 263–272.

[7] X. He, et al., "Fast matrix factorization for online recommendation with implicit feedback," in *Proc. ACM SIGIR Conf.*, 2016, pp. 549–558.

[8] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1289–1305, 2003.

[9] M. Culp and G. Michailidis, "An iterative algorithm for extending learners to a semi-supervised setting," *J. Comput. Graph. Statist.*, vol. 17, no. 3, pp. 545–571, 2008.

[10] C. Cortes and V. Vapnik, "Support vector machine," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[11] A. Dean-Hall, C. L. A. Clarke, and J. Kamps, "Overview of the trec 2012 contextual suggestion track," in *Proc. 21st Text REtrieval Conf.*, 2013, pp. 1–11.

[12] S. Gambs, M.-O. Killijian, and M. N. N. del Prado Cortez, "Next place prediction using mobility markov chains," in *Proc. 1st Workshop Meas. Privacy Mobility*, 2012, Art. no. 3.

[13] M. Xie, H. Yin, H. Wang, F. Xu, W. Chen, and S. Wang, "Learning graph-based poi embedding for location-based recommendation," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 15–24.

[14] H. Wu, J.-A. Tan, W. S. Ng, M. Xue, and W. Chen, "FTT: A system for finding and tracking tourists in public transport services," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2015, pp. 1093–1098.

[15] B. McKercher, "A chaos approach to tourism," *Tourism Manage.*, vol. 20, no. 4, pp. 425–434, 1999.

[16] N. Shoval and M. Isaacson, *Tourist Mobility and Advanced Tracking Technologies*. Evanston, IL, USA: Routledge, 2009.

[17] R. Hartmann, "Combining field methods in tourism research," *Ann. Tourism Res.*, vol. 15, no. 1, pp. 88–105, 1988.

[18] A. D. Kemperman, A. W. Borgers, and H. J. Timmermans, "Tourist shopping behavior in a historic downtown area," *Tourism Manage.*, vol. 30, no. 2, pp. 208–218, 2009.

[19] J. Connell and S. J. Page, "Exploring the spatial patterns of car-based tourist travel in loch lomond and trossachs national park, scotland," *Tourism Manage.*, vol. 29, no. 3, pp. 561–580, 2008.

[20] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated GPS traces," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, 2011, Art. no. 2.

[21] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[22] M. Xue, H. Wu, W. Chen, W. S. Ng, and G. H. Goh, "Identifying tourists from public transport commuters," in *Proc. ACM SIGKDD Conf.*, 2014, pp. 1779–1788.

[23] H. Yin and B. Cui, *Spatio-Temporal Recommendation in Social Media*. New York, NY, USA: Springer, 2016.

[24] W. Wang, H. Yin, L. Chen, Y. Sun, S. Sadiq, and X. Zhou, "Geosage: A geographical sparse additive generative model for spatial item recommendation," in *Proc. ACM SIGKDD Conf.*, 2015, pp. 1255–1264.

[25] C. Yang, L. Bai, C. Zhang, Q. Yuan, and J. Han, "Bridging collaborative filtering and semi-supervised learning: A neural approach for POI recommendation," in *Proc. ACM SIGKDD Conf.*, 2017, pp. 1245–1254.

[26] Y. Liu, T.-A. N. Pham, G. Cong, and Q. Yuan, "An experimental evaluation of point-of-interest recommendation in location-based social networks," *Proc. VLDB Endowment*, vol. 10, no. 10, pp. 1010–1021, 2017.

[27] H. Yin, X. Zhou, B. Cui, H. Wang, K. Zheng, and Q. V. H. Nguyen, "Adapting to user interest drift for POI recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2566–2581, Oct. 2016.

[28] H. Yin, B. Cui, X. Zhou, W. Wang, Z. Huang, and S. Sadiq, "Joint modeling of user check-in behaviors for real-time point-of-interest recommendation," *ACM Trans. Inf. Syst.*, vol. 35, no. 2, 2016, Art. no. 11.

[29] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, "Exploiting geographical influence for collaborative point-of-interest recommendation," in *Proc. ACM SIGIR Conf.*, 2011, pp. 325–334.

[30] Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Time-aware point-of-interest recommendation," in *Proc. ACM SIGIR Conf.*, 2013, pp. 363–372.

[31] W. Wang, H. Yin, S. Sadiq, L. Chen, M. Xie, and X. Zhou, "Spore: A sequential personalized spatial item recommender system," in *Proc. IEEE Conf. Data Eng.*, 2016, pp. 954–965.

[32] H. Yin, B. Cui, Y. Sun, Z. Hu, and L. Chen, "LCARS: A spatial item recommender system," *ACM Trans. Inf. Syst.*, vol. 32, no. 3, 2014, Art. no. 11.

[33] B. Hu and M. Ester, "Spatial topic modeling in online social media for location recommendation," in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 25–32.

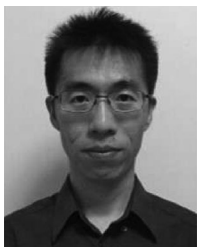
[34] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proc. ACM SIGKDD Conf.*, 2015, pp. 1235–1244.

[35] H. Yin, W. Wang, H. Wang, L. Chen, and X. Zhou, "Spatial-aware hierarchical collaborative deep learning for POI recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2537–2551, Nov. 2017.
[36] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transp. Res. Part C: Emerging Technol.*, vol. 19, no. 4, pp. 557–568, 2011.
[37] P. S. Castro, D. Zhang, C. Chen, S. Li, and G. Pan, "From taxi GPS traces to social and community dynamics: A survey," *ACM Comput. Surveys*, vol. 46, no. 2, 2013, Art. no. 17.
[38] X. Ma, C. Liu, H. Wen, Y. Wang, and Y.-J. Wu, "Understanding commuting patterns using transit smart card data," *J. Transport Geography*, vol. 58, pp. 135–145, 2017.
[39] A. Bhaskar, E. Chung, et al., "Passenger segmentation using smart card data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 3, pp. 1537–1548, Jun. 2015.
[40] J. Zhao, F. Zhang, L. Tu, C. Xu, D. Shen, C. Tian, X.-Y. Li, and Z. Li, "Estimation of passenger route choice pattern using smart card data for complex metro systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 790–801, Apr. 2017.
[41] Y. Lu, A. Misra, W. Sun, and H. Wu, "Smartphone sensing meets transport data: A collaborative framework for transportation service analytics," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 945–960, Apr. 2017.
[42] N. J. Yuan, et al., "Discovering urban functional zones using latent activity trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 712–725, Mar. 2015.
[43] Y. Lu, Z. Zeng, H. Wu, G. G. Chua, and J. Zhang, "An intelligent system for taxi service: Analysis, prediction and visualization," *AI Commun.*, vol. 31, no. 1, pp. 33–46, 2018.
[44] D. Kumar, H. Wu, Y. Lu, S. Krishnaswamy, and M. Palaniswami, "Understanding urban mobility via taxi trip clustering," in *Proc. 17th IEEE Int. Conf. Mobile Data Manage.*, 2016, vol. 1, pp. 318–324.
[45] Y. Lu, S. Xiang, and W. Wu, "Taxi queue, passenger queue or no queue?" in *Proc. 18th Int. Conf. Extending Database Technol.*, 2015, pp. 593–604.
[46] M. Qu, H. Zhu, J. Liu, G. Liu, and H. Xiong, "A cost-effective recommender system for taxi drivers," in *Proc. ACM SIGKDD Conf.*, 2014, pp. 45–54.

Yu Lu received the PhD degree in computer engineering from the National University of Singapore. He is currently an associate professor with the Faculty of Education, Beijing Normal University (BNU), where he also serves as the director of the Artificial Intelligence Lab at the Advanced Innovation Center for Future Education. His current research interests include intersection of artificial intelligence and educational technology. He has published more than 30 academic papers in prestigious journals and conferences, and currently serves as the PC member for multiple international conferences, such as the International Conference on Artificial Intelligence in Education (AIED). Before joining BNU, he was a research scientist and principle investigator with the Institute for Infocomm Research (I2R), A*STAR, Singapore.

Huayu Wu received the PhD degree in computer science from the School of Computing, National University of Singapore, in 2011. He was the head of the Data Management and Optimization Lab, Institute of Infocomm Research, A*STAR, Singapore. He has research interests in general database and data analytics topics. He is also actively engaged in applied research.

Xin Liu received the PhD degree from Nanyang Technological University, Singapore. He is currently an associate professor with the School of Automation, Hangzhou Dianzi University. Before that, he worked as a research scientist with the Institute of Infocomm Research (I2R), A*STAR, Singapore. His research interests include recommender systems, trust modeling, natural language processing, and computer vision.



Penghe Chen received the BS and PhD degrees in computer science from the National University of Singapore (NUS). He currently serves as the principle researcher with the Advanced Innovation Center for Future Education, Beijing Normal University (BNU), China. Before that, he worked with the Advanced Digital Sciences Center (ADSC), Singapore. His research interests include knowledge graph construction, data mining, and learning analytics.