Hindawi
Mobile Information Systems
Volume 2018, Article ID 1852861, 11 pages
https://doi.org/10.1155/2018/1852861

Research Article

Decision Tree-Based Contextual Location Prediction from Mobile Device Logs

Linyuan Xia, Qiumei Huang, and Dongjin Wu

School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China

Correspondence should be addressed to Qiumei Huang; hqium@mail2.sysu.edu.cn

Received 22 November 2017; Accepted 25 February 2018; Published 1 April 2018

Academic Editor: Dik Lun Lee

Contextual location prediction is an important topic in the field of personalized location recommendation in LBS (location-based services). With the advancement of mobile positioning techniques and various sensors embedded in smartphones, it is convenient to obtain massive human mobile trajectories and to derive a large amount of valuable information from geospatial big data. Extracting and recognizing personally interesting places and predicting next semantic location become a research hot spot in LBS. In this paper, we proposed an approach to predict next personally semantic place with historical visiting patterns derived from mobile device logs. To address the problems of location imprecision and lack of semantic information, a modified trip-identify method is employed to extract key visit points from GPS trajectories to a more accurate extent while semantic information are added through stay point detection and semantic places recognition. At last, a decision tree model is adopted to explore the spatial, temporal, and sequential features in contextual location prediction. To validate the effectiveness of our approach, experiments were conducted based on a trajectory collection in Guangzhou downtown area. The results verified the feasibility of our approach on contextual location prediction from continuous mobile devices logs.

1. Introduction

With the rapid development of mobile computing and positioning technology, it has made great progress in the ability and quality of location data acquisition. And nowadays, mobile phone becomes a necessity for everyone everywhere and every time which makes it convenient to capture people's daily activity trajectories. Given this, human mobility and behavior pattern analysis have become hot topics. Location-based services (LBS) has gained a great development in this few years, such as navigation services, social networking services, and personalized recommendation services. In order to provide a better service for people, it is significant to discover valuable knowledge, such as interesting locations of individuals from historical trajectories. Therefore, extracting and recognizing interesting locations and predicting next location have been an essential task for remarkable LBS. For now, despite many years of research on location prediction issue, there are still some problems: (1) using raw location data without semantic information makes it hard to study personal purpose of daily route; (2) uncleaned check-in data from social platforms increase the cost of data process and analysis despite dispersed semantic information. To deal with these problems, a contextual location prediction framework is put forward in this paper. We demonstrate the feasibility of predicting contextual location from continuous mobile devices logs by machine learning techniques. The approach proposed in this paper includes three main modules: stay point detection, semantic places recognition, and decision tree-based prediction. The first module is applied to discover individuals' behavioral sequence by extracting spatial feature from cluttered mobile device logs. Next, in order to enrich location information, the visited points extracted by the first module are attached significance in the second module through several matching methods. Based on the temporal, sequential, and semantic features of historical trajectories, a decision tree-based algorithm is applied to predict contextual location in the third module.

Regarding the prediction of people's movements, this paper attempted to predict the location type a person would

visit given his last visiting location. To our best knowledge, this has rarely been explored in the former literatures. It could be the first work to focus on contextual location prediction based on mobile device logs covering a couple of months. On the whole, this paper offers the following contributions:

(1) A modified trip-identify method is proposed to deal with consecutive mobile phone data addressing the cold-start problem for attaching location information to a more accurate extent.

(2) A semantic matching model is designed to attach place type information to the extracted stay points based on several matching rules and a self-designed POI dictionary.

(3) A decision tree-based contextual location prediction method is designed for predicting individuals' contextual location with spatial, temporal, and sequential features.

The rest of this paper is organized as follows: Section 2 introduces the related works of main procedures. Section 3 shows architecture of our approach on contextual location prediction and illustrates three processes in detail, respectively, including the modified trip-identify algorithm, semantic places recognition, and decision tree-based contextual location prediction. In Section 4, experiments are conducted to evaluate our approach. The conclusion and future work are described in Section 5.

2. Related Works

Location prediction usually refers to predicting the user's location at the next moment. Generally, two steps of stay point detection and location prediction are needed to predict locations from mobile device logs in the former researches [1–3]. In the following, previous works are reviewed with respect to the two steps.

2.1. Stay Point Detection. Pervasive location acquisition technologies produce a large amount of spatial-temporal data. Among them, GSM, WiFi, and GPS become the main sources which used in identifying people's mobility patterns [1, 2]. Call detail records (CDRs), which contains mass mobile information on call manager, is an indirect GSM-based data and allows us to reveal characteristics about the city dynamics and human behaviors [4, 5]. However, the coarse granularity, low location accuracy, and a tremendous periodic uncertainty lead to an issue of ubiquitous and continuous user-tracking capability [6]. As WiFi becomes increasingly popular in general surroundings, researchers began to discover interesting places from WiFi information by using the fingerprint-based approach [7–9]. However, the huge work on database construction makes the fingerprint-based approach unsuitable for positioning in a large-scale region. Furthermore, the outside location is hard to be obtained under the limited conditions of radio signal [10].

GPS outputs are the most common data for discovering people's visits. To find out locations where a person stays for significant time periods based on a history of successive positions, different kind of algorithms have been proposed in previous works. Clustering algorithms are popular in detecting stay points, such as K-Means clustering [11] and DJ-Cluster [12] algorithms. Besides, Palma et al. proposed a spatiotemporal clustering method, named CB-SMOT, which considers the notion of minimal time for finding clusters in single trajectories [13]. Li et al. proposed a stay point detection algorithm based on distance and pace times [14]. Zhang et al. proposed a trip-identify method in which candidate stay positions are merged in a loop until it meets a certain condition [15]. It is observed that there are different approaches aiming at different data sources and accuracy requirements. In this paper, stay points are detected following the idea of trip-identify method which can provide a fine granularity to discover individual's precise behaviors.

2.2. Location Prediction. Many researchers have paid a lot of attention on location prediction with historical human trajectories since users' mobility pattern shows a high degree of temporal and spatial regularity and hides a high degree of potential predictability despite the fact that there is individual's randomness involved [16, 17]. Based on the sequential characteristic of moving locations, studies have realized location prediction by extending the Markov model [11, 18–22]. For example, Gambs et al. proposed Mobility Markov Chain model (n-MMC) [19], and Mathew et al. trained Hidden Markov Models (HMMs) [23] for each clustered location to predict the future locations of mobile individuals. In addition to the idea based on the Markov model, machine learning is another common method in dealing with location prediction problem. It usually takes two steps in the process including pattern mining and matching. For example, Morzy proposed Traj-PrefixSpan algorithm [24], and Lei proposed a probabilistic suffix tree [25] to discover movement behaviors and predict future locations.

For better understanding human trajectories, scholars have introduced information related to location properties into trajectory analysis and put forward a concept of semantic trajectory [26]. Semantic trajectories are the track records enriched with related contextual data which give significance to meaningless raw GPS data. Considering the users' semantic-triggered intentions, Ying et al. proposed a mining-based location prediction called Geographic-Temporal-Semantic-Based Location Prediction (GTS-LP) to estimate the probability of the user in visiting a location. The main ideas of this method are to describe GTS patterns using a prefix tree and to calculate the similarity between current movements and GTS patterns by matching rules [27].

However, these previous works mainly focused on the issue of predicting users' next locations where they had been before. But, actually the predictive ability of new future location for users is more important in many cases, such as location recommendations [28]. Therefore, check-in data from social networks such as Facebook become another research data for location prediction [29, 30]. For example, Gao et al. put forward a prediction model blending social networks and the relationship of historical check-in records

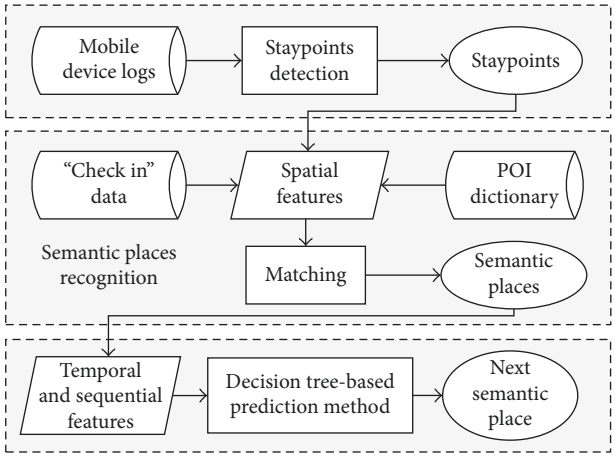

Figure 1: The workflow of the personally contextual location prediction approach.

Input: GPS, δ_{time}, δ_{distance}, δ_v
Output: Staypoints
(1) n=GPS.Count;changecount = 0;
(2) **for** i=1 ... n−1 **do**
(3) SegmentList=PointToSegment (p_i, p_{i+1});
(4) **end for**
(5) **while** changecount!=0 **do**
(6) changecount=0;
(7) SegmentList=ConcatenateSegments (SegmentList, δ_v)
(8) **for** segment **in** SegmentList **do**
(9) **if** SegmentTypeChange (segment, δ_{time}, δ_{distance}) **then**
(10) changecount+=1;
(11) **end while**
(12) **for** segment in SegmentList **do**
(13) **if** Segment.Type=='stay' **then**
(14) Staypoint=SegmentToStayPoint (Segment);
(15) **end for**

Algorithm 1: The modified trip-identify method.

of users' friends [31]. Although check-in data contain semantic information-included location properties and social relations, problems of temporal continuity and human dependency exist with this kind of data. In order to handle these problems and to achieve semantic location prediction, we propose to fuse mobile device logs, check-in data, and POIs.

3. Contextual Location Prediction

In this section, the proposed approach for predicting next contextual location is illustrated in detail. Figure 1 gives an overview of the workflow of our approach. Given the uniqueness of each individual, the GPS trajectories are manipulated and analyzed separately. In the first step, a stay point detection method is adopted to extract interesting locations from irregular and high-sampling-rate mobile device logs based on time and velocity parameters. Next, a semantic places recognition process is proposed to discover semantic information by fusing POI and check-in data. Finally, a decision tree-based method is adopted to predict next semantic place according to the extracted spatiotemporal features.

3.1. Stay Point Detection. The stay points (see Definition 1), denoting the locations where people have stayed for a while, are the most significant points in trajectories, such as restaurants for lunch and tourist attractions. To find out the stay points from mobile devices logs and improve the location accuracy, we modified the trip-identify method [15] for high-sampling-rate GPS data. The main idea of our algorithm is to determine whether the segment's type (see Definition 2) is "stay" or "move" and concatenate adjacent segments with the same state until all the neighboring segments' type are different. Algorithm of the modified trip-identify method is shown in Algorithm 1.

Definition 1. Stay point: a stay point ($\text{sp}_i = (\text{Lng}_i, \text{Lat}_i, T_{\text{arv}}, T_{\text{lev}}, \text{POI}_{\text{name}})$) stands for a location where people have stayed in a certain area for a while. T_{arv} and T_{lev} represent the timestamps that the user arrive and leave the location, respectively, and POI_{name} represents the nearest POI according to real map database.

Definition 2. Segment: a segment ($\text{segment}_i = (P_{\text{arv}}, P_{\text{lev}}, V_i, \text{type}_i)$) consists of neighboring points in time and location series with the same type. P_{arv} and P_{lev} represent two endpoints of a segment, V_i describes the velocity, and Type_i describes the state ("stay" or "move") which is determined by the corresponding velocity.

In the Algorithm 1, GPS data are turned into segments according to time series (line 2–4). For each segment in segment list, the distance and duration can be described through two endpoints as follows.

$$\begin{aligned}\text{Distance}(p_{\text{arv}}, p_{\text{lev}}) = R * \arccos(&\sin(p_{\text{arv}} \cdot \text{lat}) * \sin(p_{\text{lev}} \cdot \text{lat}) \\ &+ \cos(p_{\text{arv}} \cdot \text{lat}) * \cos(p_{\text{lev}} \cdot \text{lat}) \\ &* \cos(p_{\text{arv}} \cdot \text{lng} - p_{\text{lev}} \cdot \text{lng})) * \frac{Pi}{180}, \quad (1)\end{aligned}$$

where R represents the Earth radius (R = 6371.004 km), lat and lng represent the latitude and longitude of endpoints, respectively, and Pi is a mathematical constant and approximated as 3.14159.

$$\text{Duration}(p_{\text{arv}}, p_{\text{lev}}) = p_{\text{lev}} \cdot \text{time} - p_{\text{arv}} \cdot \text{time}, \quad (2)$$

where time attribute is the timestamp of GPS record.

The velocity and type of segments are calculated and determined based on a walking speed threshold δ_v [32] in line 7. Also, adjacent segments with the same type will be combined. If the duration of new stay segment is shorter than time threshold δ_{time}, it is considered as a move, while if the distance of move segment is less than distance threshold δ_{distance}, it is considered as a stop (line 8–10). Repeat these judgments until the segments' type wouldn't change anymore. At last, the results of stay segments are converted into stay points (line 12–15).

When dealing with GPS data, a cold-start problem should be considered to prevent missing extraction. For example, the

GPS signal would be lost immediately when people enter interior of a building, but to the contrary, GPS result would not be calculated at once when people leave the building after staying for a certain time (as shown in Figure 2). To improve location accuracy which is important in semantic place recognition process, we chose the location of arriving point as the location of stay segment instead of the center of endpoints in this case. Besides, a similar problem exists when people enter and leave from different gates of the same large building. The distance between two endpoints from different gates is greater than that from the same gate. Therefore, the distance threshold δ_{distance} should be reasonably set under the consideration of the average of building length in the experimental area.

Figure 2: An example of the effect of GPS cold-start problem.

3.2. Semantic Places Recognition. Stay points, represented in exact location with longitude and latitude, are almost meaningless in personal location description. Therefore, it is crucial to annotate stay points with location types and turn them into semantic places when taking destination as the predictive object. To endow stay points with individuals' information, we put forward a semantic places recognition process using "check in" points (see Definition 3) and POI dictionary (see Table 1).

Definition 3. "Check in" points: a "check in" point ($\text{cp}_i = (\text{Lng}_i, \text{Lat}_i, T_i, \text{type}_i)$) is similar to GPS point in expression, but the numerical values are totally dependent on the user themselves. Lng_i and Lat_i are picked up from a digital map. T_i represents the timestamp that the user checks in, and Type_i denotes the trip purpose.

To better recognize places, three steps are needed including the "check in" data matching process, clustering process, and POI dictionary matching process.

Table 1: A POI dictionary to recognize stay points without labeled information.

Type	Dictionary
Home	Dormitory, housing estate, apartment
Work	Laboratory, school, teaching building, company, office building
Restaurant	Canteen, dining room, restaurant, grogshop, hotel, pub
Shopping	Supermarket, mall, bazaar, market, department store, pedestrian street
Entertainment	Playground, cinema, KTV, swimming pool, square
Business	Hospital, administration, bank, bureau, police station, health center
Attractions	Park, museum, scenic spot, memorial hall, exhibition, temple, parkland, ruins
Others	Primary school, wharf, industrial zone, motor station, etc.

3.2.1. "Check in" Data Matching. "Check in" data are given priority compared with POI dictionary since the same physical location may imply differently for different individuals. Besides, the location types which are totally dependent on personal information can only be classified based on "check in" data such as "home" and "work." For example, a shopping mall is a place for shopping for most people, and then it falls into "shop" type. But, on the other side, it may be a place for work to salesmen, and then it should be classified as "work" type places.

As illustrated in Figure 3, the "check in" data matching rule must consider time and distance thresholds at the same time. In general, the extracted stay point matches "check in" point if the timestamp of "check in" point falls between the arriving time and leaving time of the stay point. Besides, the distance between two points is smaller than an appropriate threshold at the same time (see Definition 4).

Definition 4. Match: the matched "check in" point (cp) for a certain stay point (sp) meets the following criteria: $\text{sp}.T_{\text{arv}} < \text{cp}.T < \text{sp}.T_{\text{lev}}$ and $\text{distance}(\text{sp}, \text{cp}) < \delta_{\text{match\_distance}}$.

3.2.2. Clusters Matching. It is normal that there are some deficiencies in "check in" data because of people's forgetfulness. For most of the unmatched detected stay points, we can attach semantics to them with their address names labeled by Baidu Maps according to a self-designed POI dictionary. But in this case, it is difficult to recognize home, work, and some other places which mainly depend on individual information rather than the detailed address names. Given this, a clustering algorithm is applied to classify unmatched stay points according to the cluster type based on matched stay points. The type of each cluster is determined by the maximum probability of the corresponding stay points. The DBSCAN algorithm [33] (line 8) groups together points that are closely packed together (points with many nearby neighbors). The two required parameters, Eps and MinPts, represent the neighborhood radius and the minimum number of points to consider a point as core point, respectively. Here, it is used to gather the surrounding stay points in order to recognize places with high frequency and attach semantic type to the unmatched stay points.

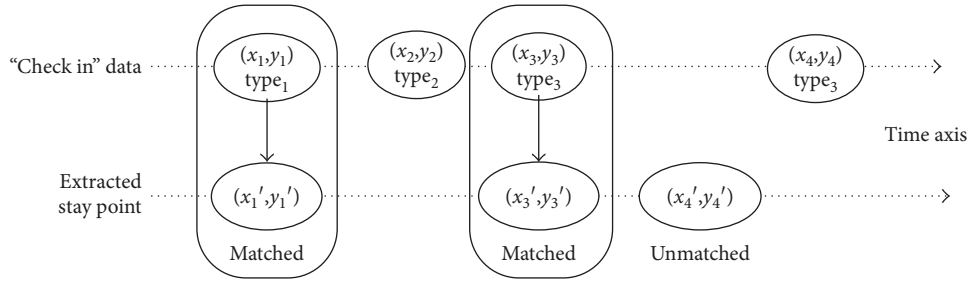

Figure 3: An example of "check in" data matching for semantic recognition.

3.2.3. POI Dictionary Matching. After matching with "check in" data, a POI dictionary matching process is applied for the still unmatched stay points. According to user survey and place category, we selected and defined several places where people stay with high frequency in daily life and designed a POI dictionary for each place type according to address name (see Table 1). The unmatched stay points will be attached with location type based on their key words of $\mathrm{POI}_{\mathrm{name}}$.

3.3. Decision Tree-Based Location Prediction. Sequential semantic trajectories of each individual are constructed after the semantic place recognition process. According to the historical movement paths, we could find the individual's activity routines and behavior patterns by decision tree method, a popular machine learning method. Usually, it takes two steps to build a decision tree, including growing a decision tree and pruning it.

3.3.1. Grow a Tree. In this paper, the ID3 decision tree algorithm [34, 35] was adopted as the main tool for executing a decision tree for our experiment. This algorithm creates a multiway tree, where there are root nodes, child nodes, branches, and leaf nodes, finding for each node the categorical feature that will yield the largest information gain for categorical targets. The formulas of information entropy and information gain are shown as follows:

$$\text{entropy}(D) = -\sum_{i=1}^{n} p_i \log_2 p_i,$$

$$\text{gain} = \text{entropy}(D) - \sum_{j=1}^{k} \frac{|D^j|}{|D|} \times \text{entropy}\left(D_j\right), \tag{3}$$

where p_i is the probability of appearance of class i in dataset D, j represents each branch node in the tree, and $|D^j|/|D|$ describes the weight of the jth partition. The attribute with highest information gain is selected to be the best extended branch for the corresponding node.

Given the people's behavior pattern in daily life, we took corresponding contextual information including temporal and sequential features into consideration in the decision tree construction (see Table 2). Temporal features include the day of week and the time of day, which represent the specific leaving time from one place to another. Sequential feature refers to the moving sequence. There is high correlation between two successive locations. For this reason, we took the present location as a sequential feature in the next place prediction. Since we assumed that all the input features are discrete values, the time attribute was divided into 24 sections. For example, "12" domain ranges from 11:30 am to 12:30 am, which is lunch time for most people. Besides, location types were predefined according to the specific places with high frequency and purpose. These specific places include the following types: Home (cover dormitory for students), Work (cover laboratory building for students), Restaurant (cover canteen for workers or students), Shopping (cover all the shopping area, from small supermarket to large shopping mall), Entertainment (cover playground, cinema, KTV, etc.), Business (cover communal services, e.g., hospital, administration, etc.), Attractions (cover park, museum, scenic spot, etc.), and Others (cover places except for the above). In this way, the temporal and spatial characteristics are easier to be utilized in feature analysis.

Table 2: The contextual information used in decision tree construction.

Feature	Description
Day of week	Mon, Tue, Wen, Thu, Fri, Sat, Sun
Time of day	1, 2, . . . , 23, 24
Present location	Home, Work, Restaurant, Shopping, Entertainment, Business, Attractions, Others

3.3.2. Prun the Tree. Decision trees are created for individuals in location type prediction according to the feature selection above. When a decision tree is built, some branches might reflect noises from the training data. Then a pruning process is carried out to solve the data over-fitting problem. ID3 uses pessimistic pruning, which makes use of error rates estimated from the training set, to replace subtree with a leaf node. This leaf is labelled with the most frequent class among the subtree being replaced [36]. An example of a decision tree structure for location prediction is shown in Figure 4. It follows a top-down approach, which starts with a training set of tuples and their associated feature labels. Based on the personalized decision tree, next location type can be calculated by providing the user's present spatial-temporal-semantic feature. For example, we can predict that the user will go for shopping when at noon at Tuesday after work. In this case, it would be incredibly helpful in places

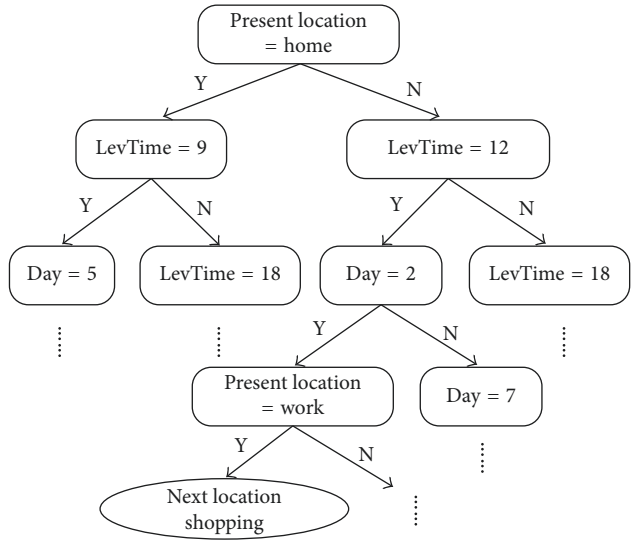

Figure 4: An example of a decision tree structure for location prediction.

Table 3: Some information of collected examples.

Items	User 1	User 2	User 3	User 4	User 5
GPS points	435822	456682	1354073	1335542	2351170
"Check in" points	468	279	322	448	147
Valid days	88	82	81	84	83
Size of data (MB)	48.1	49.2	146	133	239

recommendation system to realize location type prediction according to current geographic position.

4. Experiments and Results

In this section, we will introduce the evaluation method and give the comparative experimental analysis of the proposed contextual location prediction method.

4.1. Data Description. In this paper, we performed our experiments with two datasets: Geolife dataset [37] which is collected by Microsoft Research Asia and our own collected dataset. In the Geolife dataset, GPS trajectories are represented by a sequence of time-stamped coordinates collected by 178 users in a period of over three years from 2007 to 2011 in Beijing, China. Our own dataset is real mobility data (GPS, BDS) which are collected by 14 participants for three months (from 2016-10-15 to 2017-01-15) through their GPS-enabled smartphones. All the participants live in Guangzhou, China, and basically lead a regular life. Seven of them are office workers, and the others are students. A self-made program was running in participants' smartphones all day long for recording their daily mobility data continuously. In addition to GPS data (date, time, and coordinates), location description and POIs are also recorded through Baidu Maps API. As for "check in" data, participants are asked to check in by clicking on a digital map when they visit a place and stay more than 10 minutes. In the meantime, they should choose a place type from a predefined list based on the purpose (going for work, going for dinner, etc.) which used to improve the accuracy of places recognition in semantic places recognition process.

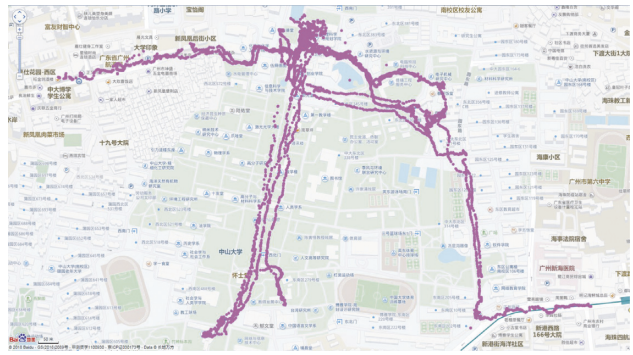

Figure 5: An example of user trajectories in one day.

We finally collected 15568306 GPS points (13561 points per person a day at average) and 5064 "check in" records (4.4 records per person a day at average). Some information of collected examples are shown in Table 3. Figure 5 shows an example of a participant's trajectories in one day.

4.2. Stay Point Detection. In order to verify the feasibility of the modified trip-identify method (MTI), we compared our method with the classical stay point detection algorithm (SPD) [14] and the original trip-identify method (OTI) [15]

Table 4: Comparisons of the stay point detection methods with the Geolife dataset.

	Stay point detection algorithm (SPD) [14]			Original trip-identify method (OTI) [15]			Modified trip-identify method (MTI)		
	100 m	200 m	500 m	100 m	200 m	500 m	100 m	200 m	500 m
Precision	0.5931	0.6192	0.6914	0.6726	0.6850	0.6927	0.6822	0.6983	0.6964
Recall	0.4947	0.5756	0.6642	0.5850	0.5914	0.6563	0.5926	0.6073	0.6616
F-measure	0.5103	0.5801	0.6644	0.6154	0.6190	0.6524	0.6239	0.6339	0.6670

using two datasets, the public dataset, and self-collected dataset. F-measure [38], which is a measure of experiment accuracy in statistical analysis, is considered as the evaluation criteria to evaluate the performances of stay point detection. The formulas of the statistics are as follows:

To investigate the performances on accuracy of different area sizes among three algorithms (SPD, OTI, and MTI), we conducted experiments with different distance values while time threshold was set as 10 minutes. The labels recorded by volunteers are used to judge the results. The experimental results are shown in Table 4. We find that these three algorithms are comparable when the distance value is 500 meters. However, when the distance value is smaller than 500 meters, that is 200 or 100 meters, it is obvious that two trip-identify methods perform better than the SPD algorithm. The F-measures of them exceed by 8.0% and 21.4% at average, respectively. It demonstrates that trip-identify methods work better in extracting stay points on a larger spatial scale.

Our own dataset is used to test the effectiveness of the MTI method focusing on a small distance again. The parameters are set as follows, as the distance value is 100 m, time threshold is 10 minutes, and the walking speed threshold is 0.5 m/s [32]. "Check in" points are treated as the true values and used to determine whether the detected stay points are correct. Table 5 presents the comparisons of performances between the proposed MTI method and the other ones. As can be seen from the table, two trip-identify methods are obviously superior to the SPD algorithm with high-sampling-rate-mobile device logs, and the MTI method is slightly better than the OTI method by solving the cold-start problem.

$$\text{precision} = \frac{\text{the correct detected stay points}}{\text{all the detected stay points}},$$

$$\text{recall} = \frac{\text{the correct detected stay points}}{\text{all the real stay points}}, \quad (4)$$

$$\text{F} - \text{measure} = \frac{2 \times \text{precison} \times \text{recall}}{\text{precison} + \text{recall}}.$$

Extensive experiments were conducted to find out a suitable distance value for the MTI method considering different types of users. Figures 6(a) and 6(b) present the recall rates and precisions with respect to different distance values. In this case, the recall rate is opposite to the distance value, the distance value is larger, and the recall rate is smaller. However, the changing law of the precision is different from the recall rate. The precision first goes up and then stays around 80% when the distance value is larger than 100 m. The number of the detected stay points is correlated with the distance value. It is known that the distance value is smaller and more stay points are detected. In the meantime, the more stay points ensure the recall rate to some degree. But, the precision is relatively low for the reason that an excess of stay points are judged as errors when the timestamps have not matched to corresponding "check in" points.

Table 5: Staypoints extraction results of three methods on our own collected data (100 m).

	SPD	OTI	MTI
Precision	0.4918	0.7685	0.7933
Recall	0.6861	0.7028	0.7208
F-measure	0.5664	0.7236	0.7445

As for the comparison between office workers and college students, both the recall rate and precision are slightly different on two types of people. For example, living in school campus, students mainly move among teaching buildings, canteens, and dormitories. Most stay points are located in a small area; thus, the recall of college students decreases more quickly than that of workers. To determine the best distance value, F-measure rate, which synthesizes the recall and precision, is taken as the main consideration. Although there is different impact on the F-measure with the same distance value between office workers and college students, the changing law of F-measure is coincident. Besides, there is no significant impact on parameter selection with two types of people. In order to simplify and unify the experiment, we chose a proper distance parameter value (80 m) for the next process for the reason that the values of F-measure (Figure 6(c)) are basically consistent when the distance parameter is set to 60 m or 80 m for both workers and students.

4.3. Semantic Place Recognition. After stay point detection, we tried to turn the coordinate positions into semantic places by using "check in" data and POI dictionary. To better obtain people's destinations of their trips, especially home and work places, the labels ("check in" data) should be given prior consideration in the matching process. To recognize special unmatched stay points, DBSCAN algorithm is adopted to find out individuals' frequent regions for the second step. At last, key word matching process with POI dictionary is also applied for the still unmatched stay points.

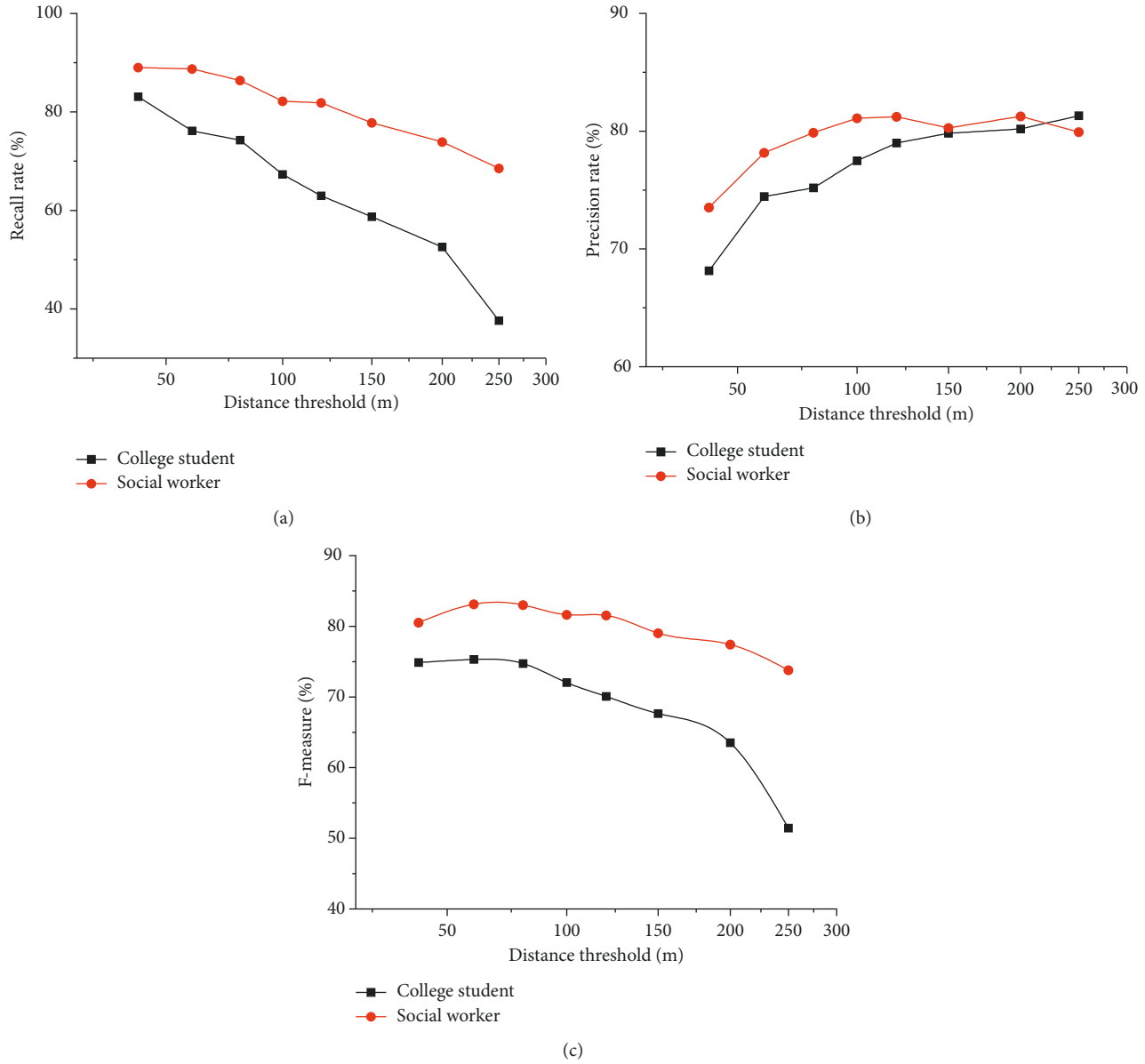

Figure 6: The performance of the algorithm based on different distance parameters. (a) The effect of distance parameters on the recall. (b) The effect of distance parameters on the precision. (c) The effect of distance parameters on the F-measure.

For the first step, a traversal method is applied to match a corresponding check-in point for each single stay point according to timestamp and position attributes. The related matching rule has been mentioned in Section 3; then we will not go into all the details of "check in" data matching processes here. During the second step, as mentioned above, the value of Eps and MinPts in DBSCAN algorithm have a great impact on individuals' frequent locations extraction and place type estimation. Thus, an experiment was conducted to find out the proper parameter values for discovering people's locations of interest.

Figure 7 shows the number of extracted clusters with different parameter values and indicates that the number of the extracted frequent locations is opposite to both parameters. The Eps and MinPts values are smaller; the number of extracted frequent locations is larger. But frequent locations approach to constant around two to three as the values of Eps and MinPts increase. Figure 8 gives an example of frequent location clusters. Considering the performance of clustering, we compared the result with the truths from participants and chose number three as the clustering number of frequent locations taking the common daily lives into consideration. According to the clustering result, the stationary point of the figure appears at (6, 40). Thus, we chose 6 and 40 as the proper MinPts and Eps values during the clusters matching process. After that, the unmatched stay points away from the frequent locations clusters are attached with semantic information based on self-designed POI dictionary as mentioned in Section 3.

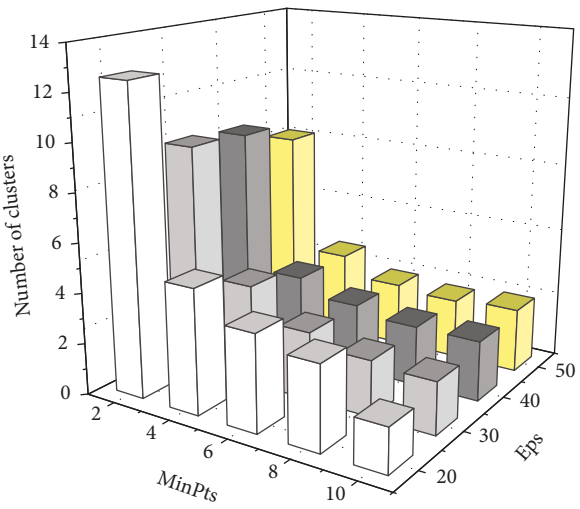

Figure 7: Performance of frequent locations extraction with different Eps and MinPts value.

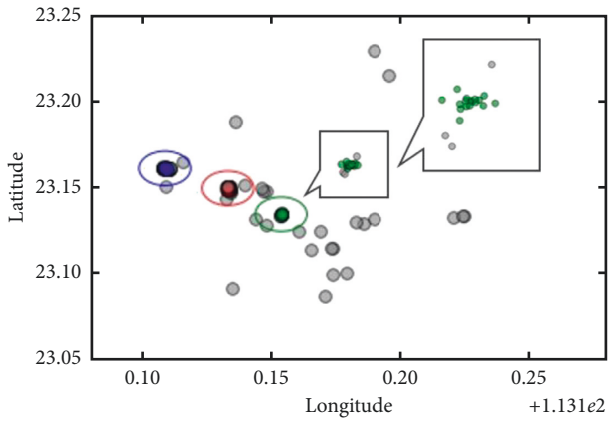

Figure 8: An example of frequent location clusters.

4.4. Decision Tree-Based Location Prediction. To validate the suitability of the decision tree model in location type prediction, the classical Markov model [19] was used for comparison. The Markov model is a stochastic model used to model randomly changing systems. It assumes that future states depend only on the current state. As for the contextual location prediction here, a set of states corresponds to the locations types extracted from spatial and semantic features. The Markov transferring matrix consists of the probabilities extracted from temporal and sequential features. They are calculated based on the number of times of each historical route. A higher Markov probability in the transferring matrix indicates that the corresponding transferring route is a more frequent route in user's daily life.

In the experiment, mobile devices logs collected in the two former months were used as training data, while the data of the last month were used as testing data. Likewise, precision, recall rate, and F-measure are used again to evaluate the performances of the prediction models. Figure 9 shows the F-measure comparisons of the contextual location type

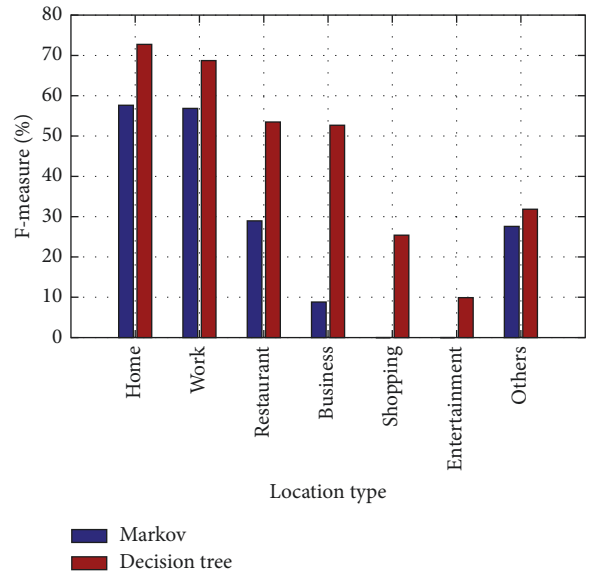

Figure 9: Performance of location type prediction based on the Markov model and the decision tree model.

prediction performance between the decision tree model and the Markov model. It is obvious that home and work places are considered as the most frequently visited locations for all the participants. This pattern is one of the main characteristics of human's daily life. As for the restaurant type, it is tested as a medium frequency visited place since people sometimes have dinner at home or company. In this case, F-measure of the Markov model is lower than that of the decision tree model. In addition to these periodic activities, occasional activities, such as business and shopping, also become predictable by using historical mobile devices logs. Since the Markov model is a probability statistic model based on maximum probability theory. Only the maximum probability is considered in the Markov model while spatial-temporal characteristics are all exploited in the decision tree model. Here, attraction type cannot be predicted by both methods. The reason is that younger generations prefer staying at home to relax rather than going outside for entertainment these years. And it can be inferred that data of three months are not so sufficient as to fail to build a perfect tree and predict attraction type. However, the above results indicate that the decision tree model achieves better contextual location prediction performance for individuals.

Figure 10 shows the experimental result of one participant. We can see from the figure that he may be an office worker who basically follows a daily routine between home and work place. Both the Markov and decision tree models perform pretty well (about 70%) on type prediction of "Home" and "Work." However, it turns out to be a problem that small probability events have usually been ignored in the Markov model. By contrast, the decision tree model is able to predict restaurants, business, and the others in spite of low recall (lower than 40%). It proves that the decision tree model has better performances especially on prediction of types with low frequency by making full use of spatial-temporal features. Overall, the prediction of commercial

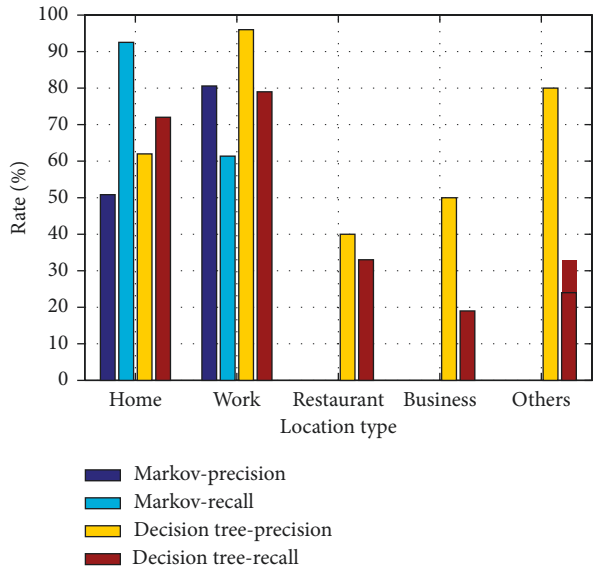

Figure 10: An example about location type prediction of one participant based on Markov and decision tee.

location type like restaurants, shopping, and entertainment can be predicted better by using the decision tree model.

5. Conclusions and Future Work

In this paper, we proposed a contextual location prediction framework for better personalized location recommendation in LBS by predicting next personally semantic place from mobile devices logs. It consists of three main modules: stay point detection, semantic places recognition, and decision tree-based prediction. The performances of each module have been evaluated with collected real-world dataset. The stay point detection results show that the modified trip-identify method extracts more precise locations with the challenge of cold-start problem compared with classical stay point detection and the original trip-identify method. A clustering algorithm and a designed POI dictionary have proven to be effective in semantic places recognition for dealing with the problem of the lack of semantic information on mobile data. The decision tree-based method, which has better performance in prediction compared with classical Markov model especially in location with low frequency, is applied for individuals' intention prediction. On the whole, the feasibility of the proposed contextual location prediction framework has been proved.

To the best of our knowledge, our work is the first to explore contextual location prediction based on the mobile devices logs collected by participants last for a couple of months. So, the proposed approach may inevitably have several limitations. For example, the prediction rate is easy to be affected by the quality of the real-world dataset. Besides, the specific value parameters in algorithms mentioned above depend on life experiences and repetitive testing. And in case office worker change their jobs or student change the schedules in a new term which has impact on personal activity prediction in real-time situation, the more recent historical trajectories can be endowed with a larger weight in training process. Based on the above considerations, we will focus on the improvement of the adaptability of this approach including parameters self-adjusting and real-time capability in the future.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Key Science and Technology Planning Project of Guangdong Province (No. 2015B010104003), the Key Science and Technology Planning Projects of Guangzhou (No. 201604046007), the National Key Research and Development Program of China (No. 2017YFB0504103), National Natural Science Foundation of China (No. 41704020), and the Fundamental Research Funds for the Central Universities (No. 17lgpy43).

References

[1] Y. Lu and Y. Liu, "Pervasive location acquisition technologies: opportunities and challenges for geospatial studies," *Computers, Environment and Urban Systems*, vol. 36, no. 2, pp. 105–108, 2012.

[2] Y. Ye, Y. Zheng, Y. Chen, J. Feng, and X. Xie, "Mining individual life pattern based on location history," in *Proceedings of the 10th International Conference on Mobile Data Management: Systems, Services and Middleware (MDM'09)*, pp. 1–10, Taipei, Taiwan, May 2009.

[3] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "WhereNext: a location predictor on trajectory pattern mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 637–646, Paris, France, June 2009.

[4] S. Isaacman, R. Becker, R. Caceres et al., "Identifying important places in people's lives from cellular network data," in *Lecture Notes in Computer Science*, vol. 6696, pp. 133–151, Springer, Berlin, Germany, 2011.

[5] M. Mamei, M. Colonna, and M. Galassi, "Automatic identification of relevant places from cellular network data," *Pervasive and Mobile Computing*, vol. 31, pp. 147–158, 2016.

[6] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Computer Networks*, vol. 64, pp. 296–307, 2014.

[7] M. Brunato and R. Battiti, "Statistical learning theory for location fingerprinting in wireless LANs," *Computer Networks*, vol. 47, no. 6, pp. 825–845, 2005.

[8] J. Hightower, S. Consolvo, A. Lamarca, I. Smith, and J. Hughes, "Learning and recognizing the places we go," in *Proceedings of the 7th International Conference on Ubiquitous Computing (UbiComp'05)*, pp. 159–176, Tokyo, Japan, September 2005.

[9] D. Kim, J. Hightower, R. Govindan, and D. Estrin, "Discovering semantically meaningful places from pervasive RF-beacons," in *Proceedings of the 11th International Conference on Ubiquitous Computing(UbiComp'09)*, pp. 21–30, Orlando, FL, USA, September 2009.

[10] M. Lv, L. Chen, Z. Xu, Y. Li, and G. Chen, "The discovery of personally semantic places based on trajectory data mining," *Neurocomputing*, vol. 173, pp. 1142–1153, 2016.

[11] D. Ashbrook and T. Starner, "Using GPS to learn significant locations and predict movement across multiple users," *Personal and Ubiquitous Computing*, vol. 7, no. 5, pp. 275–286, 2003.

[12] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personally meaningful places: an interactive clustering approach," *ACM Transactions on Information Systems*, vol. 25, no. 3, p. 12, 2007.

[13] A. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC)*, pp. 863–868, Fortaleza, Ceara, Brazil, March 2008.

[14] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma, "Mining user similarity based on location history," in *Proceedings of the 16th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems*, pp. 34:1–34:10, Irvine, CA, USA, November 2008.

[15] J. Zhang, P. Qiu, Z. Xu, and M. Du, "A method to identify trip based on the mobile phone positioning data," *Journal of Wuhan University of Technology*, vol. 37, no. 5, pp. 934–938, 2013.

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

[17] C. Song, Z. Qu, N. Blumm, and A. Barabasi, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[18] M. Chen, X. Yu, and Y. Liu, "Mining moving patterns for predicting next location," *Information Systems*, vol. 54, pp. 156–168, 2015.

[19] S. Gambs, M. Killijian, and M. Cortez, "Next place prediction using mobility Markov chains," in *Proceedings of EuroSys 2012 Workshop on Measurement, Privacy, and Mobility (MPM)*, pp. 1–6, Bern, Switzerland, April 2012.

[20] A. Asahara, K. Maruyama, A. Sato, and K. Seto, "Pedestrian-movement prediction based on mixed Markov-chain model," in *Proceedings of the 19th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems (ACM-GIS)*, pp. 25–33, Chicago, IL, USA, November 2011.

[21] W. Huang, S. Li, X. Liu, and Y. Ban, "Predicting human mobility with activity changes," *International Journal of Geographical Information Science*, vol. 29, no. 9, pp. 1569–1587, 2015.

[22] S. Cho, "Exploiting machine learning techniques for location recognition and prediction with smartphone logs," *Neurocomputing*, vol. 176, pp. 98–106, 2016.

[23] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden Markov models," in *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 911–918, Pittsburg, PA, USA, September 2012.

[24] M. Morzy, "Mining frequent trajectories of moving objects for location prediction," in *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM)*, pp. 667–680, Leipzig, Germany, July 2007.

[25] P. Lei, T. Shen, W. Peng, and I.-J. Su, "Exploring spatial-temporal trajectory model for location prediction," in *Proceedings of the 12th IEEE International Conference on Mobile Data Management (MDM)*, pp. 58–67, Luleå, Sweden, June 2011.

[26] C. Parent, S. Spaccapietra, C. Renso et al., "Semantic trajectories modeling and analysis," *ACM Computing Surveys*, vol. 45, no. 4, pp. 1–32, 2013.

[27] J. Ying, W. Lee, and V. Tseng, "Mining geographic-temporal-semantic patterns in trajectories for location prediction," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 1, pp. 1–33, 2013.

[28] J. Zhang, C. Chowmember, and Y. Li, "iGeoRec: a personalized and efficient geographical location recommendation framework," *IEEE Transactions on Services Computing*, vol. 8, no. 5, pp. 701–714, 2015.

[29] C. Jonathan and S. Eric, "Location 3: how users share and respond to location-based data on social networking sites," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pp. 74–80, Barcelona, Spain, July 2011.

[30] A. Tarasov, F. Kling, and A. Pozdnoukhov, "Prediction of user location using the radiation model and social check-ins," in *Proceedings of the 2nd ACM SIGKDD International Conference on Urban Computing*, p. 7, Chicago, USA, August 2013.

[31] H. Gao, J. Tang, X. Hu et al., "Modeling temporal effects of human mobile behavior on location-based social networks," in *Proceedings of the Conference on Information and Knowledge Management*, pp. 1673–1678, San Francisco, CA, USA, August 2013.

[32] J. Du and L. Aultmanhall, "Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: automatic trip end identification issues," *Transportation Research Part A-Policy and Practice*, vol. 41, no. 3, pp. 220–232, 2007.

[33] M. Ester, H. Kriegel, J. Sander, and X. Xiaowei, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Knowledge Discovery and Data Mining*, vol. 96, no. 34, pp. 226–231, 1996.

[34] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[35] J. S. Lee and E. S. Lee, "Exploring the usefulness of a decision tree in predicting people's locations," *Procedia-Social and Behavioral Sciences*, vol. 140, no. 4, pp. 447–451, 2014.

[36] B. B. Nair, V. P. Mohandas, and N. R. Sakthivel, "A decision tree- rough set hybrid system for stock market trend prediction," *International Journal of Computer Applications*, vol. 6, no. 9, pp. 1–6, 2010.

[37] Y. Zheng, X. Xie, and W. Ma, "Geolife: a collaborative social networking service among user, location and trajectory," *IEEE Data Engineering Bulletin*, vol. 33, no. 2, pp. 32–40, 2010.

[38] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.